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

Ruijie Zheng [1] **Yongyuan Liang** [1] **Xiyao Wang** [1] **Shuang Ma** [2]
Hal Daumé III [1] **Huazhe Xu** [3] **John Langford** [2]
Praveen Palanisamy [2] **Kalyan Shankar Basu** [2] **Furong Huang** [1]

Abstract

We present Premier-TACO, a multitask feature representation learning approach designed to improve few-shot policy learning efficiency in sequential decision-making tasks. Premier-TACO leverages a subset of multitask offline datasets for pretraining a *general feature representation*, which captures critical environmental dynamics and is fine-tuned using minimal expert demonstrations. It advances the temporal action contrastive learning (TACO) objective, known for state-of-the-art results in visual control tasks, by incorporating a novel negative example sampling strategy. This strategy is crucial in significantly boosting TACO's computational efficiency, making large-scale multitask offline pretraining feasible. Our extensive empirical evaluation in a diverse set of continuous control benchmarks including Deepmind Control Suite, MetaWorld, and LIBERO demonstrate Premier-TACO's effectiveness in pretraining visual representations, significantly enhancing few-shot imitation learning of novel tasks. Our code, pretraining data, as well as pretrained model checkpoints will be released at https://github.com/PremierTACO/premier-taco.

1. Introduction

In the dynamic and ever-changing world we inhabit, the importance of sequential decision-making (SDM) in machine learning cannot be overstated. Unlike static tasks, sequential decisions reflect the fluidity of real-world scenarios, from robotic manipulations to evolving healthcare treatments. Just as foundation models in language, such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019; Brown et al., 2020), 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 language, these models capture the essence of syntax, semantics, and context, serving as a robust starting point for a myriad of downstream tasks. Analogously, 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.

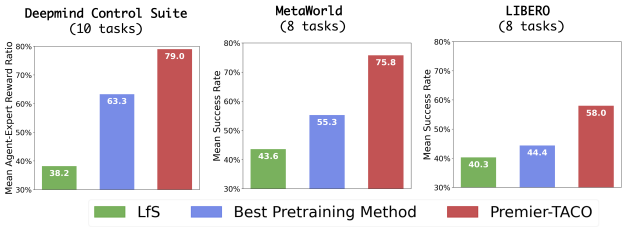

Figure 1: Performance of Premier-TACO pretrained visual representation for few-shot imitation learning on downstream unseen tasks from Deepmind Control Suite, MetaWorld, and LIBERO. LfS here represents *learning from scratch.*

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 (Lee et al., 2021). **(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**:

[1]Department of Computer Science, University of Maryland, College Park [2]Microsoft Research [3]Tsinghua University. Correspondence to: Ruijie Zheng <rzheng12@umd.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

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 (Brohan et al., 2023; Stooke et al., 2021b). Our **aspirational criteria** for foundation model for sequential decision-making encompass several key features: **(W1) Versatility** that allows the 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 Nair et al. (2022).

In light of these challenges and desirables in building foundation models for SDM, our approach to develop foundational models for sequential decision-making focuses on creating a universal and transferable encoder using a reward-free, dynamics based, temporal contrastive pretraining objective. This encoder would be tailored to manage tasks with complex observation spaces, such as visual inputs. By excluding reward signals during the pretraining stage, the model will be better poised to generalize across a broad array of downstream tasks that may have divergent objectives. Leveraging a world-model approach ensures that the encoder learns a compact representation that can capture universal transition dynamics, akin to the laws of physics, thereby making it adaptable for multiple scenarios. This encoder enables the transfer of knowledge to downstream control tasks, even when such tasks were not part of the original pretraining data set.

Existing works apply self-supervised pre-training from rich vision data such as ImageNet (Deng et al., 2009) or Ego4D datasets (Grauman et al., 2022) to build foundation models (Nair et al., 2022; Majumdar et al., 2023; Ma et al., 2023). However, applying these approaches to sequential decision-making tasks is challenging. Specifically, they often overlook control-relevant considerations and suffer from a domain gap between pre-training datasets and downstream control tasks. In this paper, rather than focusing on leveraging large vision datasets, 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. This control-centric objective learns a state representation by optimizing the mutual information between representations of current states paired with action sequences and representations of the corresponding future states.

Premier-TACO markedly enhances the effectiveness and efficiency of the temporal action contrastive learning (TACO) objective, as detailed in Zheng et al. (2023a), which delivers state-of-the-art outcomes in visual control tasks within a *single-task setting*. It extends these capabilities to efficient, large-scale *multitask offline pretraining*, broadening its applicability and performance. Specifically, while TACO considers every data point in a batch as a potential negative example, Premier-TACO strategically samples a single negative example from a proximate window of the subsequent state. This method ensures the negative example is visually akin to the positive one, necessitating that the latent representation captures control-relevant information, rather than relying on extraneous features like visual appearance. This efficient negative example sampling strategy adds no computational burden and is compatible with smaller batch sizes. In particular, on MetaWorld, using a batch size of $\frac{1}{8}$ for TACO, Premier-TACO achieves a 25% relative performance improvement. Premier-TACO can be seamlessly scaled for multitask offline pretraining, enhancing its usability and effectiveness.

Below we list our key contributions:

- ▷ **(1)** We introduce Premier-TACO, a new framework designed for the multi-task offline visual representation pretraining of sequential decision-making problems. In particular, we develop a new temporal contrastive learning objective within the Premier-TACO framework. Compared with other temporal contrastive learning objectives such as TACO, Premier-TACO employs a simple yet efficient negative example sampling strategy, making it computationally feasible for multi-task representation learning.

- ▷ **(2) [(W1) Versatility (W2) Efficiency]** Through extensive empirical evaluation, we verify the effectiveness of Premier-TACO's pretrained visual representations for few-shot learning on unseen tasks. On MetaWorld (Yu et al., 2019) and LIBERO (Liu et al., 2023), with 5 expert trajectories, Premier-TACO outperforms the best baseline pretraining method by 37% and 17% respectively. Remarkably, in LIBERO, we are the first method to demonstrate benefits from pretraining. On Deepmind Control Suite (DMC) (Tassa et al., 2018), using only 20 trajectories, which is considerably fewer demonstrations than (Sun et al., 2023; Majumdar et al., 2023), Premier-TACO achieves the best performance across 10 challenging tasks, including the hard Dog and Humanoid tasks. This versatility extends even to unseen embodiments in DMC as well as unseen tasks with unseen camera views in MetaWorld.

- ▷ **(3) [(W3) Robustness (W4) Compatability]** Furthermore, we demonstrate that Premier-TACO is not only resilient to data of lower quality but also compatible with exisiting large pretrained models. In DMC, Premier-TACO works well with the pretraining dataset collected

randomly. Additionally, we showcase the capability of the temporal contrastive learning objective of Premier-TACO to finetune a generalized visual encoder such as R3M (Nair et al., 2022), resulting in an averaged performance enhancement of around 50% across the assessed tasks on Deepmind Control Suite and MetaWorld.

2. Preliminary

2.1. Multitask Offline Pretraining

We consider a collection of tasks $\{\mathcal{T}_i : (\mathcal{X}, \mathcal{A}_i, \mathcal{P}_i, \mathcal{R}_i, \gamma)\}_{i=1}^N$ with the same dimensionality in observation space $\mathcal{X}$. Let $\phi : \mathcal{X} \to \mathcal{Z}$ 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 (Deng et al., 2009) or Ego4D (Grauman et al., 2022). 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.

2.2. TACO: Temporal Action Driven Contrastive Learning Objective

Temporal Action-driven Contrastive Learning (TACO) (Zheng et al., 2023a) 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:

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

Here, $Z_t = \phi(X_t)$ and $U_t = \psi(A_t)$ represent latent state and action variables. Theoretically, it could be shown that maximization of this mutual information objective lead to state and action representations that are capable of representing the optimal value functions. Empirically, TACO estimate the lower bound of the mutual information objective by the InfoNCE loss, and it achieves the state of 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

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 (Zheng et al., 2023a), 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 pretraining, the online learning objective of TACO (Zheng et al., 2023a) poses a notable challenge. Specifically, TACO's mechanism, which utilizes the InfoNCE (van den Oord et al., 2019) 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 as shown in Figure 2.

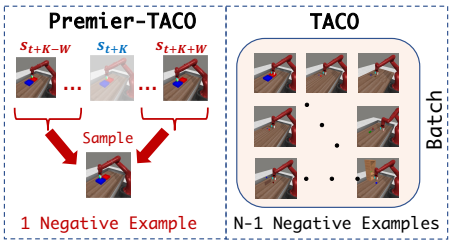

Figure 2: Difference between Premier-TACO and TACO for sampling negative examples.

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 capture temporal dynamics differentiating between positive and negative examples. In practice, this allows us to use much smaller batch sizes for Premier-TACO. On MetaWorld, with only $\frac{1}{8}$ of the batch size (512 vs. 4096), Premier-TACO achieves a 25% performance gain compared to TACO, saving around 87.5% of computational time.

In Figure 3, we illustrate the design of Premier-TACO

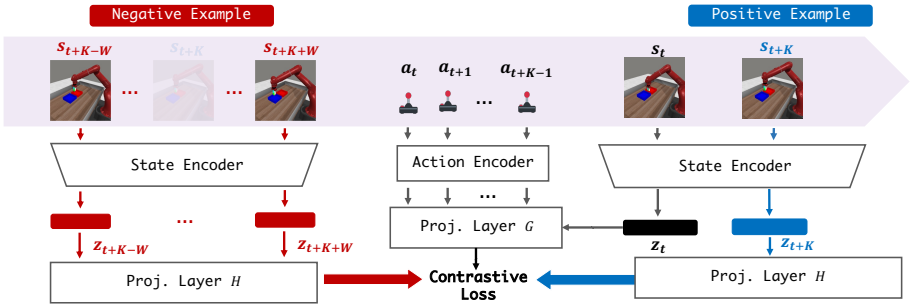

Figure 3: An illustration of Premier-TACO contrastive loss design. The two 'State Encoder's are identical, as are the two 'Proj. Layer H's. One negative example is sampled from the neighbors of framework s_{t+K}. In contrast to model-based learning, which explicitly models the transition dynamics (Janner et al., 2019; Zheng et al., 2023b; Hansen et al., 2022a; Hafner et al., 2023; Wang et al., 2024), Premier-TACO implicitly captures the environment dynamics information inside the learned latent representation.

objective. 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 $g_t^{(i)} = G_\theta(z_t^{(i)}, u_t^{(i)}, ..., u_{t+K-1}^{(i)})$, $\widetilde{h_t^{(i)}} = H_\theta(\widetilde{z_{t+K}^{(i)}})$, and $h_t^{(i)} = H_\theta(z_{t+K}^{(i)})$ as embeddings of future predicted and actual latent states. We optimize: $\mathcal{J}_{\text{Premier-TACO}}(\phi, \psi, G_\theta, H_\theta) = -\frac{1}{N}\sum_{i=1}^N \log \frac{g_t^{(i)\top} h_{t+K}^{(i)}}{g_t^{(i)\top} h_{t+K}^{(i)} + \widetilde{g_t^{(i)}}^\top h_{t+K}^{(i)}}$

Few-shot Generalization. After pretraining the representation encoder, we leverage our pretrained model Φ to learn policies for downstream tasks. To learn the policy π with the state representation $\Phi(s_t)$ as inputs, we use behavior cloning (BC) with a few expert demonstrations. For different control domains, we employ significantly fewer demonstrations for unseen tasks than what is typically used in other baselines. This underscores the substantial advantages of Premier-TACO in few-shot generalization. More details about the experiments on downstream tasks will be provided in Section 4.

4. Experiment

In our empirical evaluations, we consider three benchmarks, Deepmind Control Suite (Tassa et al., 2018) for locomotion control, MetaWorld (Yu et al., 2019) and LIBERO (Liu et al., 2023) for robotic manipulation tasks. It is important to note the varied sources of data employed for pretraining in these benchmarks. For the Deepmind Control Suite, our pretraining dataset comes from the replay buffers of online reinforcement learning (RL) agents. In MetaWorld, the dataset is generated through a pre-defined scripted policy. In LIBERO, we utilize its provided demonstration dataset, which was collected through human teleoperation. By evaluating on a wide range of pretraining data types that have been explored in previous works, we aim to provide a comprehensive evaluation for the pretraining effects of Premier-TACO.

Deepmind Control Suite (DMC): We consider a selection of 16 challenging tasks from Deepmind Control Suite. Note that compared with prior works such as (Majumdar et al., 2023; Sun et al., 2023), 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 (Yarats et al., 2022) 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. In terms of the encoder architecture, we pretrain Premier-TACO with the same shallow ConvNet encoder as in DrQv2 (Yarats et al., 2022).

MetaWorld: 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. After adding such a noise, the success rate of collected policies on average is only around 20% across ten pretrained tasks. We use the same encoder network architecture as DMC.

LIBERO: We pretrain on 90 short-horizon manipulation tasks (**LIBERO-90**) with human demonstration dataset provided by the original paper. For each task, it contains 50

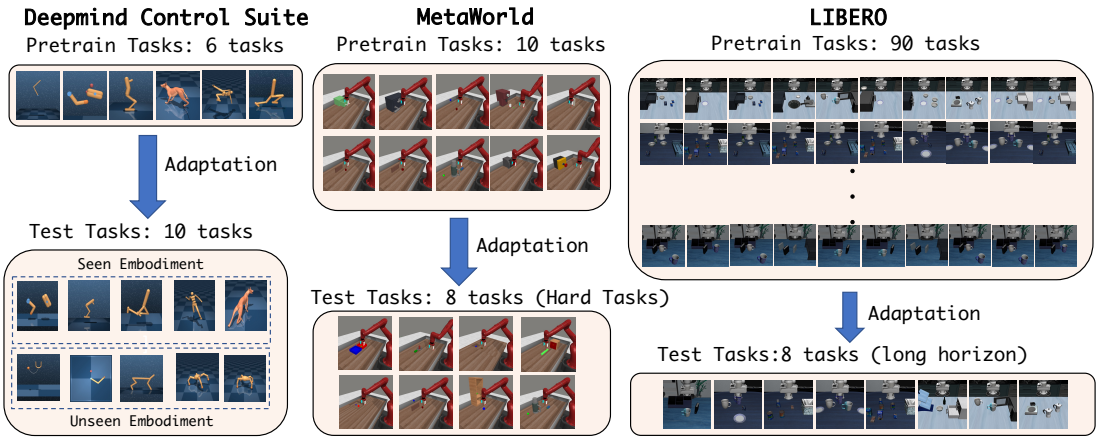

Figure 4: Pretrain and Test Tasks split for Deepmind Control Suite, MetaWorld and Libero. The left figures are Deepmind Control Suite tasks and the right figures MetaWorld tasks.

	DMControl Tasks	Models									
		LfS	SMART	Best PVRs	TD3+BC	Inverse	CURL	ATC	SPR	TACO	**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	28.4±9.7	**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	21.4±3.4	**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	30.6±6.1	**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	29.1±8.1	**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	80.1±4.1	**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	88.7±3.2	**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	58.3±6.4	**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	23.2±3.3	**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	63.9±9.3	**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	50.8±5.7	**76.8 ± 7.5**
Mean Performance		38.2	42.9	52.3	63.3	61.7	50.4	52.7	62.4	47.5	**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**.

MetaWorld Unseen Tasks	Models									
	LfS	SMART	Best PVRs	TD3+BC	Inverse	CURL	ATC	SPR	TACO	**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	67.3 ± 7.5	**78.5 ± 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	51.3 ± 10.8	**86.7 ± 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	56.8 ± 4.2	**75.0 ± 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	36.3 ± 4.5	**62.7 ± 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	52.9 ± 7.3	**72.7 ± 7.3**
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	37.8 ± 8.5	**80.2 ± 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	25.8 ± 5.0	**70.4 ± 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	52.0 ± 10.5	**80.0 ± 8.1**
Mean	43.6	53.9	53.6	51.5	51.0	48.3	51.9	55.3	47.5	**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.

trajectories of human teleoperated trajectories. We use ResNet18 encoder (He et al., 2016) to encode the image observations of resolution 128×128. For the downstream task, we assess the few-shot imitation learning performance on the first 8 long-horizon tasks of LIBERO-LONG.

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

- ▷ Learn from Scratch: Behavior Cloning with randomly initialized shallow ConvNet encoder. We carefully implement the behavior cloning from scratch baseline. For DMC and MetaWorld, following (Hansen et al., 2022b), we include the random shift data augmentation into behavior cloning. For LIBERO, we take the ResNet-T model architecture in (Liu et al., 2023), which uses a transformer decoder module on top of the ResNet encoding to extract temporal information from a sequence of observations, addressing the non-Markovian characteristics inherent in human demonstration.
- ▷ Policy Pretraining: We first train a multitask policy by TD3+BC (Fujimoto & Gu, 2021) on the pretraining dataset. While numerous alternative offline RL algorithms exist, we choose TD3+BC as a representative due to its simplicity and great empirical performance. For LIBERO, we use Multitask BC since offline RL in generally does not perform well on the imitation learning benchmark with human demonstrated dataset. 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 (Parisi et al., 2022), MVP (Xiao et al., 2022), R3M (Nair et al., 2022) and VC-1 (Majumdar et al., 2023), and report the best performance of these PVRs models for each task.
- ▷ Control Transformer: SMART (Sun et al., 2023) 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 (Laskin et al., 2020), ATC (Stooke et al., 2021a), SPR (Schwarzer et al., 2021a;b). 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 objecive (Grill et al., 2020) into sequential decision making problems by pretraining state representations that are self-predictive of future states.

Pretrained feature representation by Premier-TACO facilitates effective few-shot adaptation to unseen tasks. We measure the performance of pretrained visual representation for few-shot imitation learning of unseen downstream tasks in both DMC and MetaWorld. In particular, 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. Note that we only use $\frac{1}{5}$ of the number of expert trajectories used in (Majumdar et al., 2023) and $\frac{1}{10}$ of those used in (Sun et al., 2023).

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 and LIBERO, we use **5 expert trajectories** for all downstream tasks, and we use task success rate as the performance metric. In Table 1 Table 2, and Figure 5 we present the results for Deepmind Control Suite, MetaWorld, and LIBERO, respectively. 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. In LIBERO, consistent with what is observed in (Liu et al., 2023), existing pretraining methods on large-scale multitask offline dataset fail to enhance downstream policy learning performance. In particular, methods like multitask pretraining actually degrade downstream policy learning performance. In contrast, using ResNet-18 encoders pretrained by Premier-TACO significantly boosts few-shot imitation learning performance by a substantial margin.

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 Figure 8 (left), Premier-TACO pre-trained representation realizes an **82%** performance gain, demonstrating the robust generalizability of our pre-trained feature representations.

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 8 (right), 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 (Hansen et al., 2022c; Seo et al., 2022). Then for

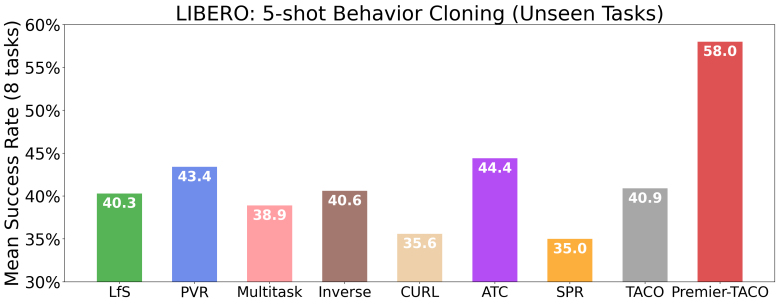

Figure 5: [(W1) Versatility (W2) Efficiency] Mean success rate of 5-shot imitation learning for 8 unseen tasks in LIBERO. Results are aggregated over 4 random seeds. **Bold** numbers indicate the best results. See the results for individual tasks in Table 4.

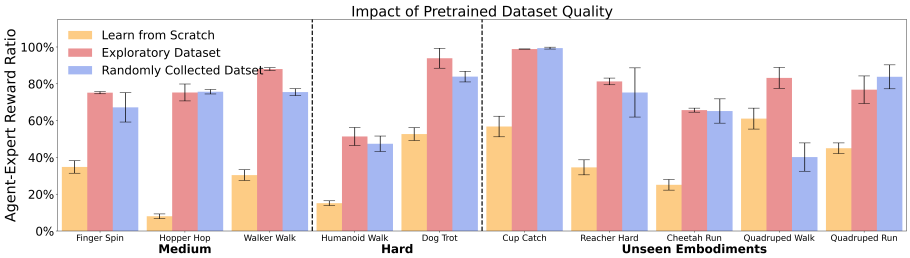

Figure 6: [(W3) Robustness] Premier-TACO pretrained with exploratory dataset vs. Premier-TACO pretrained with randomly collected dataset

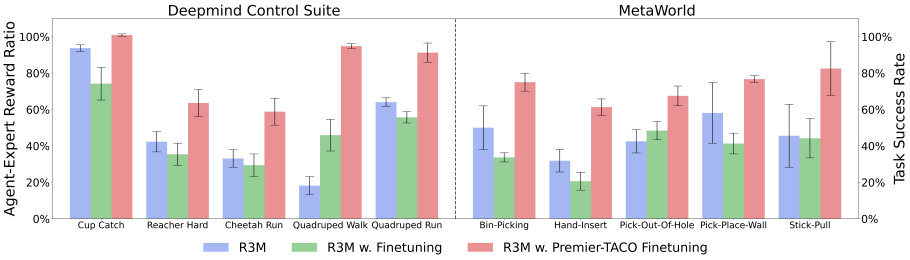

Figure 7: [(W4) Compatibility] Finetune R3M (Nair et al., 2022), a generalized Pretrained Visual Encoder with Premier-TACO learning objective vs. R3M with in-domain finetuning in Deepmind Control Suite and MetaWorld.

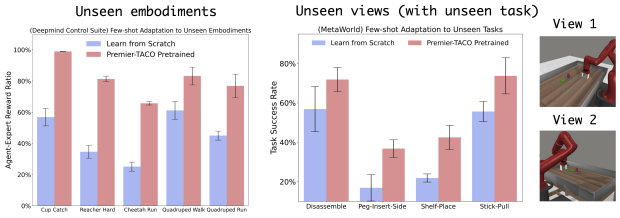

Figure 8: [(W1) Versatility] (**Left**) DMC: Generalization of Premier-TACO pre-trained visual representation to unseen embodiments. (**Right**) MetaWorld: Few-shot adaptation to unseen tasks from an unseen camera view

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.

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 6, 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.

Pretrained visual encoder finetuning with Premier-TACO. In addition to evaluating our pretrained representations across various downstream scenarios, we also conducted fine-tuning on pretrained visual representations using in-domain control trajectories following Premier-TACO framework. Importantly, our findings deviate from the observations made in prior works like (Hansen et al., 2022b) and (Majumdar et al., 2023), where fine-tuning of R3M (Nair et al., 2022) 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.

To further validate the effectiveness of our Premier-TACO approach, we compared the results of R3M with no fine-tuning, in-domain fine-tuning (Hansen et al., 2022b), and fine-tuning using our method on selected Deepmind Control Suite and MetaWorld pretraining tasks. Figure 7 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 Study - 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. This implies the promising potential to significantly improve the performance of existing pretrained models across diverse domains. The full results of finetuning on all 18 tasks including Deepmind Control Suite and MetaWorld are in Appendix B.1. 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 and TACO with different batch sizes and compare their few-shot imitation learning performance.

Figure 9 (left) 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 significantly outperform TACO across all batch sizes tested in the experiments, and exhibits performance saturation beyond a batch size of 4096. This observation substantiate that the negative example sampling strategy employed by Premier-TACO is indeed the key for the success of multitask offline pretraining.

Ablation Study - 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 preceding experiments, a consistent window size of 5 was applied across all trials on both the DeepMind Control Suite and MetaWorld. Here in Figure 9 (right) we empirically evaluate the effects of varying window sizes on the average performance of our model across ten DeepMind Control Tasks. 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.

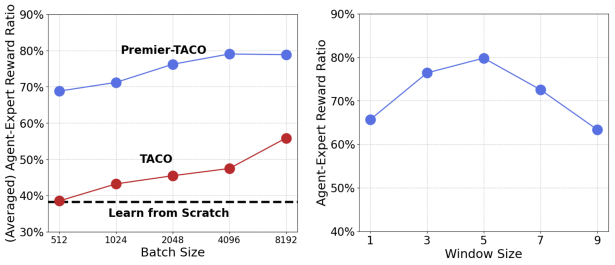

Figure 9: [(W1) Versatility] (**Left**) Premier-TACO vs. TACO on 10 Deepmind Control Suite Tasks across different batch sizes. (**Right**) Averaged performance of Premier-TACO on 10 Deepmind Control Suite Tasks across different window sizes

5. Related Work

Existing works, including R3M (Nair et al., 2022), VIP (Ma et al., 2023), MVP (Xiao et al., 2022), PIE-G (Yuan et al., 2022), and VC-1 (Majumdar et al., 2023), focus on self-supervised pre-training for building foundation models but struggle with the domain gap in sequential decision-making tasks. Recent studies, such as one by Hansen et al. (2022b), indicate that models trained from scratch often outperform pre-trained representations. Approaches like SMART (Sun et al., 2023) and DualMind (Wei et al., 2023) offer control-centric pre-training, but at the cost of extensive fine-tuning or task sets. Contrastive learning techniques like CURL (Laskin et al., 2020), CPC (Henaff, 2020), ST-DIM (Anand et al., 2019), and ATC (Stooke et al., 2021a) have succeeded in visual RL, but mainly focus on high-level features and temporal dynamics without a holistic consideration of state-action interactions, a gap partially filled by TACO (Zheng et al., 2023a). Our work builds upon these efforts but eliminates the need for extensive task sets and fine-tuning, efficiently capturing control-relevant features. This positions our method as a distinct advancement over DRIML (Mazoure et al., 2020) and Homer (Misra et al., 2019), which require more computational or empirical resources.

A detailed discussion of related work is in Appendix A.

6. Conclusion

This paper introduces Premier-TACO, a robust and highly generalizable representation pretraining framework for few-shot policy learning. We propose a temporal contrastive learning objective that excels in multi-task representation learning during the pretraining phase, thanks to its efficient negative example sampling strategy. Extensive empirical evaluations spanning diverse domains and tasks underscore the remarkable effectiveness and adaptability of Premier-TACO's pre-trained visual representations to unseen tasks, even when confronted with unseen embodiments, different views, and data imperfections. Furthermore, we demonstrate the versatility of Premier-TACO by showcasing its ability to fine-tune large pretrained visual representations like R3M (Nair et al., 2022) with domain-specific data, underscoring its potential for broader applications.

Acknowledgements

Zheng, Wang, and Huang are supported by National Science Foundation NSF-IIS-2147276 FAI, DOD-ONR-Office of Naval Research under award number N00014-22-1-2335, DOD-AFOSR-Air Force Office of Scientific Research under award number FA9550-23-1-0048, DOD-DARPA-Defense Advanced Research Projects Agency Guaranteeing AI Robustness against Deception (GARD) HR00112020007, Adobe, Capital One and JP Morgan faculty fellowships.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ajay, A., Kumar, A., Agrawal, P., Levine, S., and Nachum, O. OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *ICLR*, 2021. URL https://openreview.net/forum?id=V69LGwJ0lIN. 15

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.-A., and Hjelm, R. D. Unsupervised state representation learning in atari. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6fb52e71b837628ac16539c1ff911667-Paper.pdf. 8, 14

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.-H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-1: Robotics transformer for real-world control at scale, 2023. 2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. 1

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf. 15

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 3

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. 1, 18

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20132–20145. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a8166da05c5a094f7dc03724b41886e5-Paper.pdf. 6

Gao, Y., Zhang, R., Guo, J., Wu, F., Yi, Q., Peng, S., Lan, S., Chen, R., Du, Z., Hu, X., Guo, Q., Li, L., and Chen, Y. Context shift reduction for offline meta-reinforcement learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 80024–80043. Curran Associates, Inc., 2023. 15

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S. K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E. Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P. R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G. M., Fuegen, C., Ghanem, B., Ithapu, V. K., Jawahar, C. V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H. S., Rehg, J. M., Sato, Y., Shi, J., Shou, M. Z., Torralba, A., Torresani, L., Yan, M., and Malik, J. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. 2, 3

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf. 6

Gupta, A., Kumar, V., Lynch, C., Levine, S., and Hausman, K. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *CoRL*, 2019. URL https://proceedings.mlr.press/v100/gupta20a.html. 15

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS. 14

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models, 2023. 4

Hansen, N., Wang, X., and Su, H. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022a. 4

Hansen, N., Yuan, Z., Ze, Y., Mu, T., Rajeswaran, A., Su, H., Xu, H., and Wang, X. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. In *CoRL 2022 Workshop on Pre-training Robot Learning*, 2022b. URL https://openreview.net/forum?id=tntIAuQ50E. 6, 7, 8, 14, 15

Hansen, N. A., Su, H., and Wang, X. Temporal difference learning for model predictive control. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 17–23 Jul 2022c. URL https://proceedings.mlr.press/v162/hansen22a.html. 6, 14

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. 5

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4182–4192. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/henaff20a.html. 8, 14

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019. 4

Ji, T., Liang, Y., Zeng, Y., Luo, Y., Xu, G., Guo, J., Zheng, R., Huang, F., Sun, F., and Xu, H. Ace : Off-policy actor-critic with causality-aware entropy regularization, 2024. 14

Jiang, Z., Zhang, T., Janner, M., Li, Y., Rocktäschel, T., Grefenstette, E., and Tian, Y. Efficient planning in a compact latent action space. In *ICLR*, 2023. URL https://openreview.net/forum?id=cA77NrVEuqn. 15

Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., and Larlus, D. Hard negative mixing for contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21798–21809. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7cade80b7cc92b991cf4d2806d6bd78-Paper.pdf. 14

Kim, M., Rho, K., Kim, Y.-d., and Jung, K. Action-driven contrastive representation for reinforcement learning. *PLOS ONE*, 17(3):1–14, 03 2022. doi: 10.1371/journal.pone.0265456. URL https://doi.org/10.1371/journal.pone.0265456. 14

Kipf, T., Li, Y., Dai, H., Zambaldi, V., Sanchez-Gonzalez, A., Grefenstette, E., Kohli, P., and Battaglia, P. Compile: Compositional imitation learning and execution. In *ICML*, 2019. 15

Laskin, M., Srinivas, A., and Abbeel, P. CURL: Contrastive unsupervised representations for reinforcement learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5639–5650. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/laskin20a.html. 6, 8, 14

Lee, S., Seo, Y., Lee, K., Abbeel, P., and Shin, J. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=AlJXhEI6J5W. 1

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 14

Liu, B., Zhu, Y., Gao, C., Feng, Y., qiang liu, Zhu, Y., and Stone, P. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=xzEtNSuDJk. 2, 4, 6, 17, 18, 19

Ma, S., Zeng, Z., McDuff, D., and Song, Y. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OMizHuea_HB. 14

Ma, Y. J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., and Zhang, A. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YJ7o2wetJ2. 2, 8, 14

Majumdar, A., Yadav, K., Arnaud, S., Ma, Y. J., Chen, C., Silwal, S., Jain, A., Berges, V.-P., Abbeel, P., Malik, J., Batra, D., Lin, Y., Maksymets, O., Rajeswaran, A., and Meier, F. Where are we in the search for an artificial visual cortex for embodied intelligence?, 2023. 2, 4, 6, 7, 8, 14, 16

Mazoure, B., Tachet des Combes, R., Doan, T. L., Bachman, P., and Hjelm, R. D. Deep reinforcement and infomax learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3686–3698. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/26588e932c7ccfa1df309280702fe1b5-Paper.pdf. 8, 14

Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. Discovering and achieving goals via world models. *Advances in Neural Information Processing Systems*, 34:24379–24391, 2021. 14

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. *CoRR*, abs/1911.05815, 2019. URL http://arxiv.org/abs/1911.05815. 8, 14

Mitchell, E., Rafailov, R., Peng, X. B., Levine, S., and Finn, C. Offline meta-reinforcement learning with advantage weighting, 2021. URL https://openreview.net/forum?id=S5S3eTEmouw. 15

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., and Gupta, A. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=tGbpgz6yOrI. 2, 3, 6, 7, 8, 9, 14, 15, 16

Nam, H. and Han, B. Learning multi-domain convolutional neural networks for visual tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4293–4302, 2016. doi: 10.1109/CVPR.2016.465. 14

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., and Lin, D. Libra r-cnn: Towards balanced learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 14

Parisi, S., Rajeswaran, A., Purushwalkam, S., and Gupta, A. The unsurprising effectiveness of pre-trained vision models for control. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17359–17371. PMLR, 17–23 Jul 2022. 6, 16

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 18

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019. 1

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CR1XOQ0UTh-. 14

Schwarzer, M., Anand, A., Goel, R., Hjelm, R. D., Courville, A., and Bachman, P. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=uCQfPZwRaUu. 6

Schwarzer, M., Rajkumar, N., Noukhovitch, M., Anand, A., Charlin, L., Hjelm, R. D., Bachman, P., and Courville, A. Pretraining representations for data-efficient reinforcement learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=XpSAvlvnMa. 6

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pp. 8583–8592. PMLR, 2020. 14

Seo, Y., Hafner, D., Liu, H., Liu, F., James, S., Lee, K., and Abbeel, P. Masked world models for visual control. In *CoRL*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1332–1344. PMLR, 2022. 6

Shrivastava, A., Gupta, A., and Girshick, R. Training Region-based Object Detectors with Online Hard Example Mining. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 14

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. Decoupling representation learning from reinforcement learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9870–9879. PMLR, 18–24 Jul 2021a. 6, 8, 14

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*, pp. 9870–9879. PMLR, 2021b. 2

Sun, Y., Zheng, R., Wang, X., Cohen, A. E., and Huang, F. Transfer RL across observation feature spaces via model-based regularization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=7KdAoOsI81C. 14

Sun, Y., Ma, S., Madaan, R., Bonatti, R., Huang, F., and Kapoor, A. SMART: Self-supervised multi-task pretraining with control transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9piH3Hg8QEf. 2, 4, 6, 8, 14

Tabassum, A., Wahed, M., Eldardiry, H., and Lourentzou, I. Hard negative sampling strategies for contrastive representation learning, 2022. 14

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. Deepmind control suite, 2018. 2, 4

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding, 2019. 3, 14

Wan, S., Chen, Z., Zhang, T., Zhang, B., and kat Wong, K. Bootstrapping face detection with hard negative examples, 2016. 14

Wang, X., Zheng, R., Sun, Y., Jia, R., Wongkamjan, W., Xu, H., and Huang, F. COPlanner: Plan to roll out conservatively but to explore optimistically for model-based RL. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jnFcKjtUPN. 4

Wei, Y., Sun, Y., Zheng, R., Vemprala, S., Bonatti, R., Chen, S., Madaan, R., Ba, Z., Kapoor, A., and Ma, S. Is imitation all you need? generalized decision-making with dual-phase training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16221–16231, October 2023. 8, 14

Xiao, T., Radosavovic, I., Darrell, T., and Malik, J. Masked visual pre-training for motor control, 2022. 6, 8, 14, 16

Xu, G., Zheng, R., Liang, Y., Wang, X., Yuan, Z., Ji, T., Luo, Y., Liu, X., Yuan, J., Hua, P., Li, S., Ze, Y., III, H. D., Huang, F., and Xu, H. Drm: Mastering visual reinforcement learning through dormant ratio minimization.

In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MSe8YFbhUE. 14

Xu, M., Lu, Y., Shen, Y., Zhang, S., Zhao, D., and Gan, C. Hyper-decision transformer for efficient online policy adaptation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AatUEvC-Wjv. 15

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pp. 11920–11931. PMLR, 2021a. 14

Yarats, D., Kostrikov, I., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=GY6-6sTvGaf. 14

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_SJ-_yyes8. 4, 14, 17

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019. URL https://arxiv.org/abs/1910.10897. 2, 4

Yuan, Z., Xue, Z., Yuan, B., Wang, X., WU, Y., Gao, Y., and Xu, H. Pre-trained image encoder for generalizable visual reinforcement learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13022–13037. Curran Associates, Inc., 2022. 8, 14

Zhang, A., McAllister, R. T., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-2FCwDKRREu. 14

Zheng, R., Wang, X., Sun, Y., Ma, S., Zhao, J., Xu, H., III, H. D., and Huang, F. TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=ezCsMOy1w9. 2, 3, 8, 14

Zheng, R., Wang, X., Xu, H., and Huang, F. Is model ensemble necessary? model-based RL via a single model with lipschitz regularized value function. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=hNyJBk3CwR. 4

Zheng, R., Cheng, C.-A., III, H. D., Huang, F., and Kolobov, A. PRISE: Learning temporal action abstractions as a sequence compression problem. In *Forty-first International Conference on Machine Learning*, 2024. 15

A. Detailed Discussion of 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 (Nair et al., 2022), VIP (Ma et al., 2023), MVP (Xiao et al., 2022), PIE-G (Yuan et al., 2022), and VC-1 (Majumdar et al., 2023) 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 (Hansen et al., 2022b) 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 (Sun et al., 2023) 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 (Wei et al., 2023) pretrains representations using 45 tasks for general-purpose decision-making without task-specific fine-tuning. Besides, some methods (Sekar et al., 2020; Mendonca et al., 2021; Yarats et al., 2021a; Sun et al., 2022) 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 Visual RL (Yarats et al., 2021b; 2022; Hafner et al., 2020; Hansen et al., 2022c; Xu et al., 2024; Ji et al., 2024) is long-standing challenge due to the entangled problem of representation learning and credit assignment. 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 (Laskin et al., 2020) extracts high-level features by utilizing InfoNCE(van den Oord et al., 2019) to maximize agreement between augmented observations, although it does not explicitly consider temporal relationships between states. Several approaches, such as CPC (Henaff, 2020), ST-DIM (Anand et al., 2019), and ATC (Stooke et al., 2021a) , 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 (Mazoure et al., 2020) 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 Kim et al. (2022); Zhang et al. (2021) incorporate actions into the contrastive loss, emphasizing behavioral similarity. TACO (Zheng et al., 2023a) 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 (Misra et al., 2019) 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.

Hard Negative Sampling Strategy in Contrastive Learning Our proposed negative example sampling strategy in Premier-TACO is closely related to hard negative example mining in the literature of self-supervised learning as well as other areas of machine learning. Hard negative mining is indeed used in a variety of tasks, such as facial recognition (Wan et al., 2016), object detection (Shrivastava et al., 2016), tracking (Nam & Han, 2016), and image-text retrieval (Pang et al., 2019; Li et al., 2021), by introducing negative examples that are more difficult than randomly chosen ones to improve the performance of models. Within the regime of self-supervised learning, different negative example sampling strategies have also been discussed both empirically and theoretically to improve the quality of pretrained representation. In particular, (Robinson et al., 2021) modifies the original NCE objective by developing a distribution over negative examples, which prioritizes pairs with currently similar representations. (Kalantidis et al., 2020) suggests to mix hard negative examples within the latent space. (Ma et al., 2021) introduce a method to actively sample uncertain negatives by calculating the gradients of the loss function relative to the model's most confident predictions. Furthermore, (Tabassum et al., 2022) samples negatives that combine the objectives of identifying model-uncertain negatives, selecting negatives close to the anchor point in the latent embedding space, and ensuring representativeness within the sample population.

Comparison with Offline Meta-RL Methods Compared with offline meta-rl methods, feature representation learning with self-supervised/contrastive objectives, such as Premier-TACO, can efficiently leverage low-quality datasets (e.g., datasets collected by rolling out random actions in the DeepMind Control Suite) (Xu et al., 2023; Gao et al., 2023; Mitchell et al., 2021). Offline meta RL, in contrast, relies on datasets with good coverage to learn effective policies and typically addresses tasks with smaller shifts between meta-training and meta-testing (e.g., varying velocities in MuJoCo's halfcheetah).

In particular, HDT (Xu et al., 2023) utilizes a hyper-network as an adaptation module to encode expert demonstrations and augment the base Decision Transformer (Chen et al., 2021) model. Unlike HDT, Premier-TACO can adapt to unseen embodiments with different action spaces by initializing a new policy head. In contrast, HDT's hyper-network architecture does not easily accommodate unseen action spaces without significant modifications. CSRO (Gao et al., 2023) focuses on smaller task distribution shift as in prior offline meta RL works, such as humanoid in MuJoCo running in different directions. The task representation learned in CSRO is also not able to handle the unseen action spaces. On the contrary, Premier-TACO tackles a broader problem and experimental setting, enabling our representation to generalize to unseen downstream tasks and unseen embodiments with new action spaces.

Other pretraining schemes for decision-making. In this work, we primarily focuses on pretraining visual state representations. Several other existing works aim to solve multitask pretraining for sequential decision making from a different perspective, through the discovery of temporal action abstractions (i.e. skills or options). These works propose to pretrain temporally extended action primitives and subsequently use them for shortening the effective decision-making horizon during high-level policy induction, including CompILE (Kipf et al., 2019), RPL (Gupta et al., 2019), OPAL (Ajay et al., 2021), LOVE (Jiang et al., 2023), and PRISE (Zheng et al., 2024). They often operate in two stages: learning the primitives during the first and applying them to solve a downstream task during the second, possibly adapting the primitives in the process. Compared with the visual state representation proposed in Premier-TACO, temporal action abstractions go in an orthogonal direction, and combining the benefits of both pretrained state representation and temporal action abstractions could be an exciting future direction.

B. Additional Experiment Results

B.1. Finetuning

Comparisons among R3M (Nair et al., 2022), R3M with in-domain finetuning (Hansen et al., 2022b) and R3M finetuned with Premier-TACO in Deepmind Control Suite and MetaWorld are presented in Figure 11 and 10.

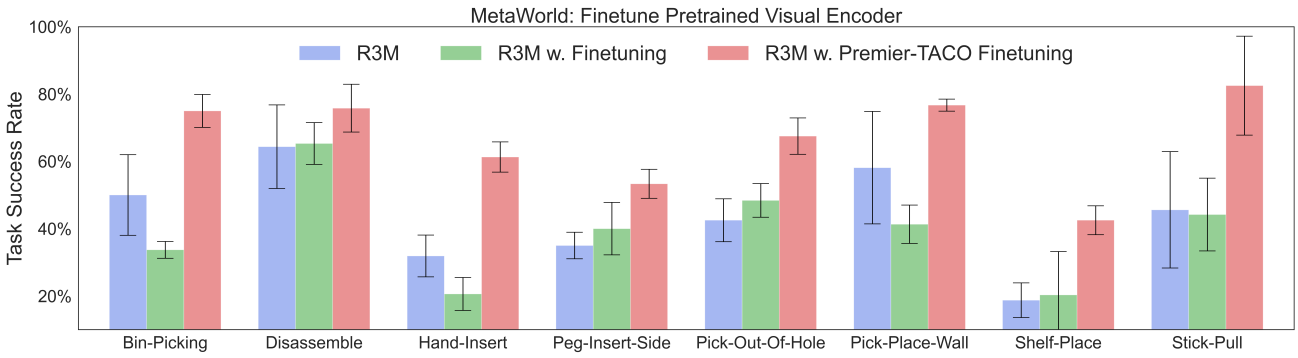

Figure 10: [(W4) Compatibility] Finetune R3M (Nair et al., 2022), a generalized Pretrained Visual Encoder with Premier-TACO learning objective vs. R3M with in-domain finetuning in Deepmind Control Suite and MetaWorld.

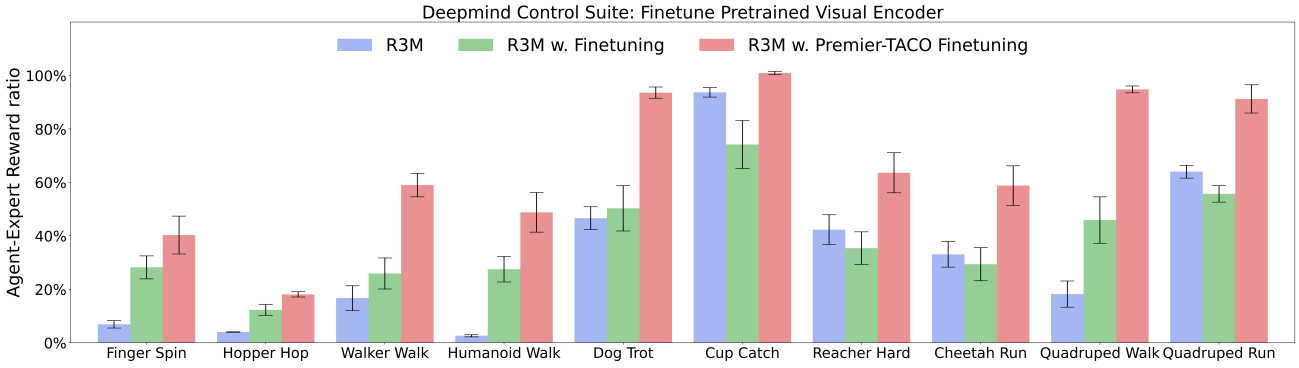

Figure 11: [(W4) Compatibility] Finetune R3M (Nair et al., 2022), a generalized Pretrained Visual Encoder with Premier-TACO learning objective vs. R3M with in-domain finetuning in Deepmind Control Suite and MetaWorld.

B.2. Pretrained Visual Representations

Here, we provide the full results for all pretrained visual encoders across all 18 tasks on Deepmind Control Suite and MetaWorld.

DMControl	Pretrained Visual Models				**MetaWorld**	Pretrained Visual Models			
	PVR	MVP	R3M	VC-1		PVR	MVP	R3M	VC-1
Finger Spin	11.5 ± 6.0	5.4 ± 7.1	6.9 ± 1.4	38.4 ± 9.3	Bin Picking	45.6 ± 5.6	46.1 ± 3.1	50.0 ± 12.0	60.2 ± 4.3
Hopper Hop	10.2 ± 1.5	7.8 ± 2.7	4.0 ± 0.1	23.2 ± 4.9	Disassemble	47.6 ± 5.8	32.4 ± 5.1	64.4 ± 12.4	70.4 ± 8.9
Walker Walk	10.3 ± 3.8	8.30 ± 1.6	16.7 ± 4.6	30.5 ± 6.2	Hand Insert	18.8 ± 4.0	10.4 ± 5.6	31.8 ± 6.21	35.5 ± 2.3
Humanoid Walk	7.6 ± 3.4	3.2 ± 0.5	2.6 ± 0.4	30.1 ± 7.5	Peg Insert Side	25.3 ± 10.4	28.9 ± 5.4	35.0 ± 3.95	48.2 ± 3.6
Dog Trot	20.5 ± 12.4	32.9 ± 6.0	46.6 ± 4.3	73.5 ± 6.4	Pick Out Of Hole	28.4 ± 5.7	42.3 ± 9.7	42.5 ± 6.4	66.3 ± 7.2
Cup Catch	60.2 ± 10.3	56.7 ± 8.9	93.7 ± 1.8	89.2 ± 13.2	Pick Place Wall	30.7 ± 8.5	42.5 ± 10.9	58.1 ± 16.7	63.2 ± 9.8
Reacher Hard	33.9 ± 9.2	40.7 ± 8.5	42.3 ± 5.6	64.9 ± 5.8	Shelf Place	19.5 ± 6.4	21.2 ± 8.3	18.7 ± 5.15	32.4 ± 6.5
Cheetah Run	26.7 ± 3.8	27.3 ± 4.4	33.1 ± 4.8	39.5 ± 9.7	Stick Pull	30.2 ± 4.6	28.5 ± 9.6	45.6 ± 17.3	52.4 ± 5.6
Quadruped Walk	15.6 ± 9.0	14.5 ± 7.2	18.2 ± 4.9	63.2 ± 4.0					
Quadruped Run	40.6 ± 6.7	43.2 ± 4.2	64.0 ± 2.4	61.3 ± 8.5					

Table 3: Few-shot results for pretrained visual representations (Parisi et al., 2022; Xiao et al., 2022; Nair et al., 2022; Majumdar et al., 2023)

B.3. LIBERO-10 success rate

LIBERO	Models								
Unseen Tasks	LfS	Best PVRs	Multitask	Inverse	CURL	ATC	SPR	TACO	**Premier-TACO**
Task 0	19.9 ± 4.1	21.2 ± 3.5	23.3 ± 4.3	23.3 ± 4.3	16.7 ± 6.2	23.8 ± 6.9	15.0 ± 0.0	5.0 ± 0.0	$\mathbf{35.5 \pm 7.5}$
Task 1	40.0 ± 8.8	46.7 ± 6.2	48.3 ± 10.3	38.3 ± 9.2	26.7 ± 8.5	41.3 ± 7.5	35.0 ± 6.3	40.3 ± 4.1	$\mathbf{70.0 \pm 5.0}$
Task 2	63.3 ± 6.2	65.8 ± 6.7	60.0 ± 4.1	51.6 ± 4.7	35.0 ± 4.1	65.0 ± 9.3	35.0 ± 5.0	65.0 ± 9.3	$\mathbf{95.0 \pm 7.2}$
Task 3	55.7 ± 4.7	56.4 ± 3.2	66.7 ± 8.4	70.1 ± 7.1	70.0 ± 6.8	$\mathbf{83.8 \pm 6.2}$	55.0 ± 5.4	62.5 ± 7.3	75.0 ± 13.2
Task 4	$\mathbf{43.3 \pm 6.2}$	27.9 ± 3.9	26.7 ± 3.1	28.3 ± 2.3	18.3 ± 2.5	25.0 ± 6.1	23.7 ± 2.2	15.5 ± 2.4	30.7 ± 2.5
Task 5	66.7 ± 9.2	62.8 ± 9.3	46.7 ± 3.8	63.3 ± 13.1	78.3 ± 2.4	78.8 ± 7.3	68.7 ± 11.9	52.3 ± 6.2	$\mathbf{80.0 \pm 6.1}$
Task 6	6.7 ± 6.2	14.5 ± 6.7	21.7 ± 2.3	11.6 ± 4.7	23.3 ± 2.4	11.2 ± 4.1	12.5 ± 5.6	19.8 ± 3.8	$\mathbf{27.5 \pm 7.2}$
Task 7	26.7 ± 4.7	29.6 ± 8.9	35.0 ± 7.1	38.3 ± 5.5	16.7 ± 2.4	26.3 ± 4.1	35.0 ± 9.3	22.3 ± 7.9	$\mathbf{50.3 \pm 4.0}$
Mean	40.3	43.4	38.9	40.6	35.6	44.4	35.0	40.9	**58.0**

Table 4: [(W1) Versatility (W2) Efficiency] Five-shot Behavior Cloning (BC) for unseen task of LIBERO. Success rate of Premier-TACO and baselines across first 8 tasks on LIBERO-10. Results are aggregated over 4 random seeds. **Bold** numbers indicate the best results.

C. Additional Experiment Results on Downstream Online Reinforcement Learning

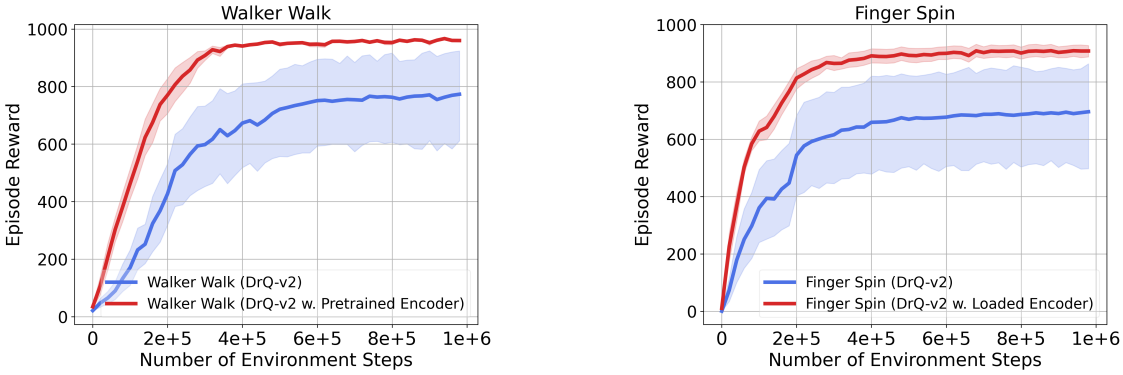

Figure 12: Downstream RL instead of imitation learning on Walker Walk (**Left**) and Finger Spin (**Right**). Results are aggregated over 8 random seeds.

While our paper primarily focuses on the sample-efficient imitation learning for downstream adaptation, we can also apply reinforcement learning instead of imitation learning. Here in Figure 12, we have included additional experimental results on two unseen tasks, Walker Walk and Finger Spin, to showcase the downstream RL performance. We choose DrQ-v2 (Yarats et al., 2022) as the backbone visual RL algorithm and compare the performance of DrQ-v2 from scratch with DrQ-v2 using pretrained Premier-TACO encoder. Notably, representation learned from pretrained Premier-TACO encoder can also significantly accelerate downstream RL learning.

D. Implementation Details

Dataset For six pretraining tasks of Deepmind Control Suite, we train visual RL agents for individual tasks with DrQ-v2 (Yarats et al., 2022) 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 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. For LIBERO, we take the human demonstration dataset from Liu et al. (2023), which contains 50 demosntration trajectories per task.

Pretraining For the shallow convolutional network used in Deepmind Control Suite and MetaWorld, we follow the same architecture as in Yarats et al. (2022) 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.

```python
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 LIBERO, we use two randomly initialized (or pretrained) ResNet-18 encoders to encode the third-person view and first-person view images with FiLM (Perez et al., 2018) encoding method to incorporate the BERT embedding (Devlin et al., 2019) of the task language instruction. During downstream behavior cloning, we apply a transformer decoder module with context length 10 on top of the ResNet encodings to extract the temporal information, and then attach a two-layer MLP with hidden size 1024 as the policy head. The architecture follows ResNet-T in Liu et al. (2023).

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. 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 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 for Deepmind Control Suite and Metaworld, and a batch size of 64 is used for LIBERO. During behavior cloning, we finetune the Shallow ConvNet Encoder. However, when we applied Premier-TACO for the large pre-trained ResNet/ViT encoder models, 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 and LIBERO, we execute 40 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 on Deepmind Control Suite and Metaworld, and RTX A5000 on LIBERO.

E. An Additional Ablation Study on Negative Example Sampling Strategy

In Premier-TACO, we sample one negative example from a size W window centered at the positive example for each data point. However, in principle, we could also use all samples within this window as negative examples instead of sampling only one. In the table below, we compare the performance of two negative example sampling strategies across 10 unseen Deepmind Control Suite tasks. **Bold numbers** indicate the better results.

	Sampling 1	**Sampling All**
Finger Spin	$\mathbf{75.2 \pm 0.6}$	70.2 ± 8.4
Hopper Hop	75.3 ± 4.6	$\mathbf{76.1 \pm 3.0}$
Walker Walk	88.0 ± 0.8	$\mathbf{88.5 \pm 0.4}$
Humanoid Walk	51.4 ± 4.9	$\mathbf{56.4 \pm 8.9}$
Dog Trot	$\mathbf{93.9 \pm 5.4}$	92.1 ± 4.0
Cup Catch	$\mathbf{98.9 \pm 0.1}$	98.3 ± 1.6
Reacher Hard	$\mathbf{81.3 \pm 1.8}$	80.1 ± 5.8
Cheetah Run	65.7 ± 1.1	$\mathbf{69.3 \pm 2.3}$
Quadruped Walk	83.2 ± 5.7	$\mathbf{85.4 \pm 4.2}$
Quadruped Run	76.8 ± 7.5	$\mathbf{82.1 \pm 9.1}$
Overall	79.0	**79.8**

Table 5: Results of two different negative sampling strategies on 10 unseen Deepmind Control Suite Tasks.

As shown in Table 5, we find that using all samples from the size W window does not significantly enhance performance compared to Premier-TACO. Moreover, this approach considerably increases the computational overhead. Given these results, we chose a more computationally efficient strategy of sampling a single negative example from the size W window in Premier-TACO.

F. Task instructions of downstream LIBERO tasks

Here in table 6, we provide the language instruction for each of the LIBERO downstream task. We refer readers to (Liu et al., 2023) for more details of the tasks.

TASK ID	TASK SCENE	TASK INSTRUCTION
0	LIVING ROOM SCENE2	PUT BOTH THE ALPHABET SOUP AND THE TOMATO SAUCE IN THE BASKET
1	LIVING ROOM SCENE2	PUT BOTH THE CREAM CHEESE BOX AND THE BUTTER IN THE BASKET
2	KITCHEN SCENE3	TURN ON THE STOVE AND PUT THE MOKA POT ON IT
3	KITCHEN SCENE4	PUT THE BLACK BOWL IN THE BOTTOM DRAWER OF THE CABINET AND CLOSE IT
4	LIVING ROOM SCENE5	PUT THE WHITE MUG ON THE LEFT PLATE AND PUT THE YELLOW AND WHITE MUG ON THE RIGHT PLATE
5	STUDY SCENE1	PICK UP THE BOOK AND PLACE IT IN THE BACK COMPARTMENT OF THE CADDY
6	LIVING ROOM SCENE6	PUT THE WHITE MUG ON THE PLATE AND PUT THE CHOCOLATE PUDDING TO THE RIGHT OF THE PLATE
7	LIVING ROOM SCENE1	PUT BOTH THE ALPHABET SOUP AND THE CREAM CHEESE BOX IN THE BASKET

Table 6: Language instructions for 8 LIBERO downstream tasks.