DisPIM: Distilling PreTrained Image Models for Generalizable Visuo-Motor Control

Haitao Wang[1,2] , **Hejun Wu**[*1,2]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, Guangdong, China
wuhejun@mail.sysu.edu.cn

Abstract

We introduce DisPIM, a framework that leverages pretrained image models (PIMs) for visuo-motor control. Applying PIMs to visuo-motor control faces a big difficulty due to the distribution shift between the distribution of visual environmental states and that of the pretraining datasets. Due to such a distribution shift, fine-tuning PIMs specifically for visuo-motor control may hurt the generalizability of PIMs, while adding additional tunable parameters for specific actions apparently leads to high computational costs. DisPIM addresses these challenges using a novel feature distillation approach, which obtains a compact model that not only inherits the generalization capability of PIMs but also acquires task-specific skills for visuo-motor control. This good for both sides is mainly achieved by means of a target Q-ensemble mechanism, which is inspired by double Q-learning. This Q-ensemble mechanism can adaptively adjust the distillation rate, so as to balance the objective of generalization and task-specific ability during training. With this balancing mechanism, DisPIM achieves both task-specific and generalizable control requiring a low computation cost. Across a series of algorithms, task domains, and evaluation metrics in both simulation and a real robot, our DisPIM demonstrates significant improvements in generalization and overall performance with low computational overhead.

1 Introduction

Visuo-motor control, which involves training a robot to make decisions based on visual inputs, may benefit from large-scale pretrained image models (PIMs). Recently, PIMs such as SAM [Kirillov *et al.*, 2023] and MAE [He *et al.*, 2022], have shown promise in performing visual control tasks in a zero-shot manner. Inspired by their success, researchers have begun applying PIMs to visuo-motor control by processing observation frames, given that these models were pretrained on out-of-domain datasets. Compared to Learning-From-Scratch (LFS) methods (Figure 1 (a)), Learning-From-PIMs (LFP) (Figure 1 (b)) can effectively enhance the generalization capability of visuo-motor control.

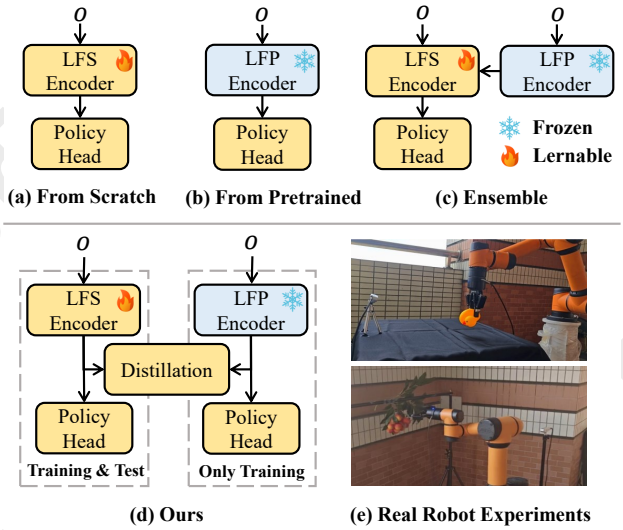


Figure 1: Pipeline comparison between (a) Learning from Scratch (LFS), (b) Learning from Pretraining (LFP), (c) Ensemble, (d) Our distillation framework, and (e) The experiments of real robot in our work

Nevertheless, directly applying PIMs to visuo-motor control faces great challenges, primarily due to the distribution shift between the visual states in environments and the large-scale datasets in pretraining [Yang *et al.*, 2023]. While fine-tuning PIMs on the specific control task seems like a possible good solution, recent studies have shown that such fine-tuning often leads to a significant loss of the model's pretrained generalization abilities [Yuan *et al.*, 2022]. An alternative solution is to employ ensemble models (Figure 1 (c)), which combine LFS and LFP approaches to harness their respective strengths and potentially mitigate distributional shift issues [Lin *et al.*, 2023]. As illustrated in Figure 2, the ensemble method outperforms LFS, LFP, and fine-tuned PIMs in the PixMC generalization benchmark [Xiao *et al.*, 2022]. Unfortunately, ensemble models introduce additional complex-

*Corresponding author

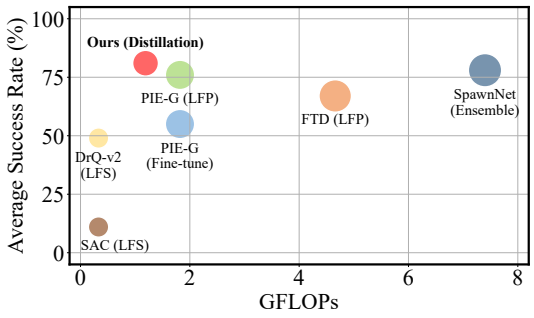


Figure 2: Comparison of generalization capability on the PixMC benchmark [Xiao *et al.*, 2022]. Bubble size indicates the number of parameters in the models. A higher number of GFLOPs implies greater computational demands during the test phase. Our method achieves a better balance between generalization capability and computational cost.

ity and computational overhead (e.g., higher GFLOPs), which can be prohibitive in resource-constrained environments.

In contrast to the trend of using larger PIMs or complex ensemble models, our study turns attention to downsizing the model while keeping their generalization capability in visuo-motor control. This leads to a critical question: *How can we effectively transfer the generalization capability of existing PIMs to a small one, thereby retaining the generalization capability with minimal computational cost?* Drawing inspiration from transfer learning in computer vision [Wu *et al.*, 2022], this work attempts to answer the question from the perspective of feature-based distillation [Romero *et al.*, 2014]. Feature-based distillation enables a small LFS model to mimic the output of a PIM, thereby allowing a small model to acquire similar generalization capability. However, applying feature-based distillation directly to visuo-motor control often results in unstable training, as shown in Figure 5. The distillation process acts as a regularization term. If the small model is pushed too closely to the PIM, it may compromise task-specific learning needed for visual control tasks. Conversely, emphasizing task-specific learning too much might lose the generalization capacity, a challenge we refer to as "*Unbalanced Feature Distillation*."

To address these challenges, we propose DisPIM, a novel feature distillation framework for visuo-motor control. DisPIM enables a control model to be both task-specific and generalizable while requiring low computational costs during the test phase. As shown in Figure 1 (d), "task-specific" is achieved through end-to-end training with a policy head, while the "generalizable" capability is achieved by using a PIM as a teacher to transfer pretrained knowledge to the smaller model.

The key innovation of our DisPIM is the Q-dynamic feature distillation mechanism. Inspired by double Q-learning that reduces the overestimation bias by using two independent Q-functions, the Q-dynamic mechanism employs a Q-ensemble module. This module dynamically balances generalization and task-specific learning. It adjusts the distillation rate, allowing the LFS encoder to maintain the generalization abilities of the LFP encoder without sacrificing task-specific knowledge. Thus, the issue of "Unbalanced Feature Distillation" is resolved. Additionally, the Q-ensemble module can be combined with upper-confidence bound (UCB) exploration to further enhance the agent's performance.

We thoroughly evaluate the performance of our DisPIM framework on three widely used generalization benchmarks: DMC-GB [Hansen and Wang, 2020], DrawerWorld [Wang *et al.*, 2021], and PixMC [Xiao *et al.*, 2022]. **We also perform real-world experiments in Aubo i5 robot.** [1] In all evaluations, our method demonstrates superior performance, showcasing its effectiveness and versatility. We highlight the contributions of this paper as follows:

1. We propose DisPIM, an effective framework for generalizable visuo-motor control. It effectively enables an control model to be both task-specific and generalizable under a small computational cost.
2. We introduce a Q-dynamic feature distillation approach to solve the "*Unbalanced Feature Distillation*" problem, ensuring that the model simultaneously achieves both task-specific and generalizable capability.
3. We perform extensive evaluations of different methods both in simulation and a real robot, showing significant improvement of our method in generalization capability.

2 Related Work

2.1 Pretrained Image Models for Generalizable visuo-motor control

Using pretrained image models (PIMs) for visuo-motor control has achieved promising results [Xiao *et al.*, 2022; Yen-Chen *et al.*, 2020; Shridhar *et al.*, 2022; Khandelwal *et al.*, 2022; Ebert *et al.*, 2021]. For example, RRL [Shah and Kumar, 2021] and PIE-G [Yuan *et al.*, 2022] use pretrained ResNet [He *et al.*, 2015] for state representation learning that can generalize to unseen visual scenarios in a zero-shot manner. VRL3 [Wang *et al.*, 2022] proposes a multi-stage pretrained framework for solving visual control tasks. APV [Seo *et al.*, 2022] introduces a framework that learns representations useful for understanding the dynamics via generative pre-training on videos. Some studies [Banino *et al.*, 2021; Du *et al.*, 2023] investigate the use of pretrained visual-language models such as BERT [Devlin *et al.*, 2018] for understanding and interpreting visual scenarios. Other works [Chen *et al.*, 2024; Wang *et al.*, 2023] leverage the promptable segmentation of SAM [Kirillov *et al.*, 2023] to enhance the generalization capabilities. However, most of the above methods require additional computational costs during fine-tuning or directly using the ensemble model. Our proposed DisPIM leverages LFP and the LFS model during the training phase, while only using the LFS model for testing, which has a low computational cost.

2.2 Feature-based Knowledge Distillation

Knowledge distillation is a technique for transferring knowledge from a teacher model to a student model. Mainstream

[1] https://www.aubo-robotics.cn/

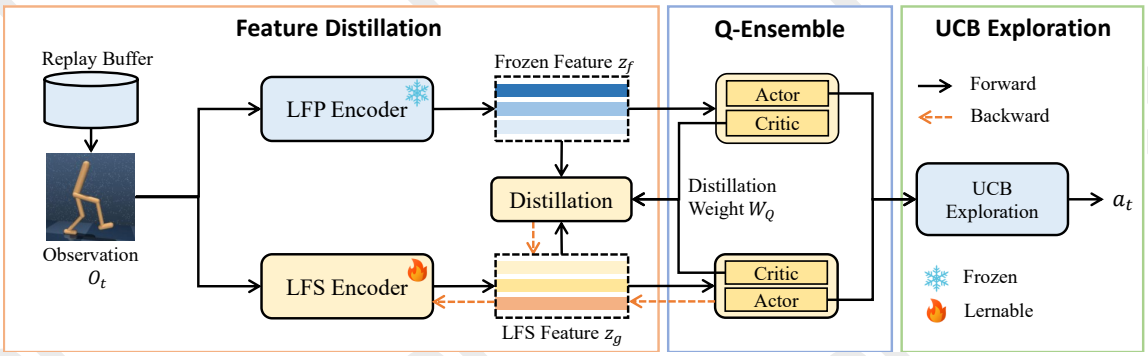


Figure 3: Overview of our DisPIM framework.

distillation methods can be broadly classified into three categories: logits-based [Hinton *et al.*, 2015], similarity-based [Tung and Mori, 2019], and feature-based [Romero *et al.*, 2014]. Research has indicated that feature-based distillation offers clearer optimization targets and surpasses the performance of the other two methods [Chen *et al.*, 2022a].

Several studies have explored distilling PIMs into more compact models through feature-based knowledge distillation. For instance, CLIPPING [Pei *et al.*, 2023] emphasizes model compression, aiming to comprehensively transfer knowledge from a large Clip4clip model [Luo *et al.*, 2021] to MobileViT-v2 [Mehta and Rastegari, 2022]. Meanwhile, VLKD [Dai *et al.*, 2022] focuses on aligning the features of the language model with the CLIP model and then integrating them into a multi-modal generator. All of these methods have the common goal of distilling knowledge from one model to another. Differently, in our case, we have two distinct objectives: preserving the generalization capability of a PIM and ensuring effective task-specific capability.

To balance these two objectives, we do not simply use a feature-based distillation method. Instead, we carefully designed a feature distillation method for visuo-motor control tasks based on a Q-ensemble. Our DisPIM allows for the flexible adaptation of features from PIMs to the LFS model without sacrificing task-specific capability.

3 Background

3.1 Visuo-Motor Control with MDP

We formulate the problem of visuo-motor control as a Markov Decision Process (MDP) [Bellman, 1957], which is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function of system that defines a probability distribution over the next state given the current state and action, where $\Delta(\mathcal{S})$ is the distribution over the state space $\mathcal{S}$. The reward function $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ assigns a scalar reward to each state-action pair (s_t, a_t), and $\gamma \in [0, 1)$ is the discount factor. The objective is to learn a policy π that maximizes the expected discounted sum of rewards:

$$L = \mathbb{E}_{s_0, a_0, s_1, a_1, \cdots} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{1}$$

where s_0 is the initial state and $a_t = \pi(s_t)$ is the action chosen by the policy at timestep t.

3.2 Feature-based Distillation

The feature-based distillation is achieved by enforcing the feature consistency between the teacher and student models:

$$L_{FD} = \frac{1}{N} \sum_{N}^{i} \| f^{(t)} - f^{(s)} \|_2, \tag{2}$$

where $f^{(t)}$, $f^{(s)}$ and N denote the teacher model, student model, and batch size, respectively. The objective of our DisPIM is to improve the generalization capability in the LFS model, which is crucial for visuo-motor control tasks.

3.3 Generalization

Given a training environment $\mathcal{M}_n$ and a set of testing environments $\mathcal{E}_{test} = \{\mathcal{M}_{n+1}, \mathcal{M}_{n+2}, \ldots, \mathcal{M}_{n+m}\}$, where each environment $\mathcal{M}_i$ is defined as a MDP represented by a tuple $(\mathcal{S}_i, \mathcal{A}_i, P_i, R_i, \gamma_i)$. The objective of generalizable visuo-motor control is to learn a control policy $\pi^* = \arg\max_\pi \mathbb{E}_{\mathcal{M}_n \sim p_{train}(\mathcal{M})}[J_{\mathcal{M}_i}(\pi)]$ in the training phase that generalizes well across both the training and testing environments. The policy π^* should perform well in training environments (task-specific), while also performing well in the unseen testing environments (generalizable).

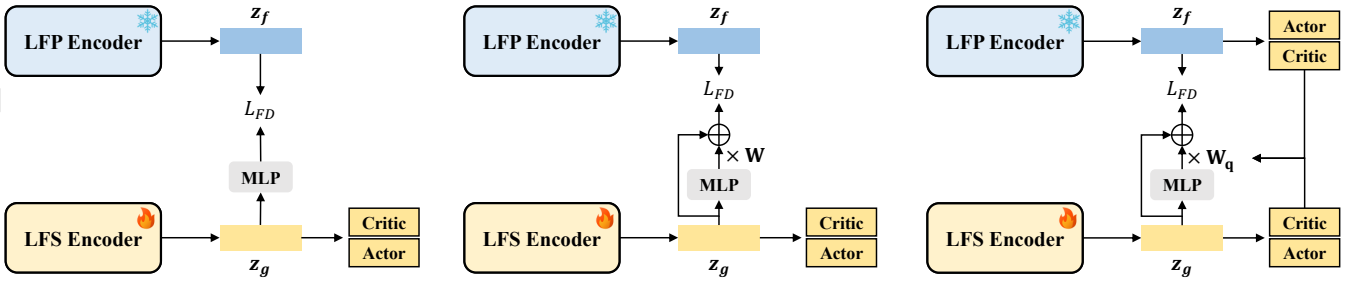
4 Method

In this section, we first provide a brief overview of the overall workflow of DisPIM. Then, we introduce our design motivations and specific implementations. Finally, we integrate an effective exploration technique into our DisPIM for exploration.

4.1 Method Overview

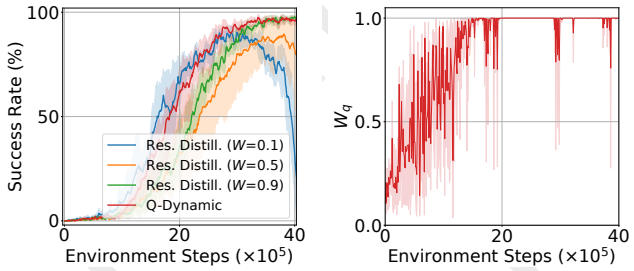
To achieve the above goal, we now present the overview of our DisPIM, as illustrated in Figure 3. In the **Training stage**, DisPIM uses three primary components: (1) Feature Distillation Module, (2) Q-ensemble Module, and (3) Exploration Module.

At each time step t, the agent receives a pixel observation o_t. Initially, we employ an LFP encoder f and an LFS encoder g to acquire the frozen and learnable feature representation $z_f^t = f(o_t) \in \mathbb{R}^D$ and $z_g^t = g(o_t) \in \mathbb{R}^D$, respectively.



(a) Feature Distillation with projector (b) Feature Distillation with Residual (c) Feature Distillation with Q-Dynamic

Figure 4: Illustration of different feature distillation approaches. (a) Using an MLP to map features from the student space to the teacher space [Chen *et al.*, 2022b]. (b) Employing a modified residual network on the student model to balance feature learning using a hyperparameter [Huang *et al.*, 2024]. (c) The proposed Q-dynamic feature distillation employs a Q-ensemble to achieve dynamic balance in feature learning. Our method aims to simultaneously optimize generalization capability and task-specific knowledge, enhancing the overall performance of the model.



(a) (b)

Figure 5: (a) The learning curves of fixed distillation rate methods and Q-dynamic method. (b) The curves of W_q.

Subsequently, these two representations are fed into two randomly initialized actor-critics to make decisions, respectively. At the same time, we take the LFP encoder as a teacher and use Q-ensemble for feature distillation. Finally, we use a ucb-based technique for exploration. In the **Test stage**, only the LFS encoder is used to visuo-motor control. Next, we present our DisPIM in detail.

4.2 Q-Dynamic Feature Distillation for control

Using an LFP encoder as the teacher model, the two common methods for implementing feature-based distillation are illustrated in Figure 4 (a) [Chen *et al.*, 2022b] and (b) [Huang *et al.*, 2024]. The method in Figure 4 (a) is to apply a projector to map z_g from the student space to the teacher space. This way expects the projection MLP(z_g) to remain the same as the pretrained ones z_f, while relaxing the constraints on z_g for fitting the control tasks. However, under such a condition, the distillation loss $L_{FD} = \|z_f - MLP(z_g)\|_2$ would be too loose for z_g to keep close with z_f, thereby limiting its generalization capability.

The other common way (as shown in Figure 4 (b)) is to use a residual network for balancing the two learning objectives when conducting distillation. This design allows the LFS features to effectively receive supervision from generalized ones while also keeping task-specific. As shown in Figure 4 (b), this design applies a residual network h on z_g to transform its representation with an MLP projector and an identity mapping:

$$h(z_g) = z_g + W \times MLP(z_g) \tag{3}$$

where W is a hyperparameter that is used to balance the learning of generalization and task-specific capability. In this way, the generalizable target z_f can directly guide the generalizable learning of z_g but not enforce z_g to be the same as z_f, which makes it flexible for z_g to fit the control tasks.

Nevertheless, we found that when incorporating the above residual distillation network structure for visuo-motor control tasks, the balancing hyperparameter W plays a crucial role in training stability. As shown in Figure 5, when the residual weight W is set to a small number ($W = 0.1$), the embedding space of z_g is largely constrained by the teacher model, causing rapid convergence in the early stages but a sharp decline later on. Conversely, with a large W ($W = 0.9$), z_g may overfit the training task, reducing generalization. Although the training curve converges smoothly, the convergence rate slows down. Based on these observations, we demonstrate that the control performance can be significantly penalized if W is not dynamically adjusted, as different training stages require different values for optimal stability and learning progression. Therefore, an intuitive idea is to relate the distillation ratio W to control performance.

Similar to how Double Q-learning addresses the overestimation of Q-values by using multiple Q-functions, we consider dynamically calculating W through a Q-ensemble mechanism. As shown in Figure 4 (c), we apply an ensemble of two control agents $\{Q_{\theta_i}, \pi_{\phi_i}\}_{i=1}^2$, where θ_i and ϕ_i denote the parameters of the i-th Q-function and policy, respectively. Each Q-function has a unique target Q-function $Q_{\bar{\theta}_i}$. For each agent i, we consider a Q-dynamic weight based on an ensemble of target Q-functions:

$$W_q(s,a) = 2\sigma\left(-\bar{Q}_{std}(s,a)\right) \tag{4}$$

where σ is the sigmoid function and $\bar{Q}_{std}(s,a)$ is the

empirical standard deviation of all target Q-functions [2] $\{Q_{\bar{\theta}_i}\}_{i=1}^{2}$. Note that the $W_q(s,a)$ is bounded in $[0,1]$, because $\bar{Q}_{std}(s,a)$ is always positive.

By dynamically linking the distillation ratio to the evaluation of state-action pairs through the Q-functions, the model optimizes the knowledge distillation process based on the actual performance of each state-action pair during training, avoiding the issues caused by a fixed distillation ratio.

After obtaining the transformed vector $\bar{z}_g = h(z_g) = z_g + W_q \times MLP(z_g)$, we can use the LFP model as the teacher to distill knowledge. The distillation loss can be written as:

$$L_{FD} = \sum_{i}^{N} \|f(o_t) - h(g(o_t))\|_2 \tag{5}$$

The overall learning objective is then the combination of reinforcement learning and distillation losses:

$$L = L_{RL} + L_{FD} \tag{6}$$

Where L_{RL} is a certain reinforcement learning algorithm loss, such as PPO [Schulman *et al.*, 2017], DrQ-v2 [Yarats *et al.*, 2021].

4.3 UCB Exploration

Since the ensemble can express higher uncertainty on unseen samples, it can also be leveraged for efficient exploration [Osband *et al.*, 2016], as shown in Figure 3. We consider an optimism-based exploration that chooses the action by the following:

$$a_t = \max_a \{Q_{mean}(s_t,a) + \lambda Q_{std}(s_t,a)\} \tag{7}$$

where Q_{mean} and Q_{std} are the empirical mean and standard deviation of all parameterized Q-functions, and λ is a hyperparameter. This inference method was originally proposed in [Osband *et al.*, 2016] for efficient exploration in DQN [Mnih *et al.*, 2013]. In this work, we extend it to our ensemble learning framework for better exploration.

4.4 Implementation

For the teacher encoder, we use the ViT-Small Encoder [Dosovitskiy *et al.*, 2020] with a 16×16 patch size, 384 hidden sizes, 6 attention heads, and 12 blocks. We use the MAE framework [He *et al.*, 2022] to pretrain the teacher model on the ImageNet dataset [Deng *et al.*, 2009]. For the student encoder, we use a 6-layer transformer encoder, with the other parameters being the same as those of the teacher encoder.

5 Experiments

5.1 Benchmark

DMC-GB [Hansen and Wang, 2020]. We evaluate the robustness in terms of the visual background changes on DMC-GB. Models are trained in an original DMControl environment [Tassa *et al.*, 2018], and we measure generalization to environments with natural videos as background. This setting consists of more complicated and fast-switching video backgrounds that are drastically different from the training environments, as shown in Figure 6.

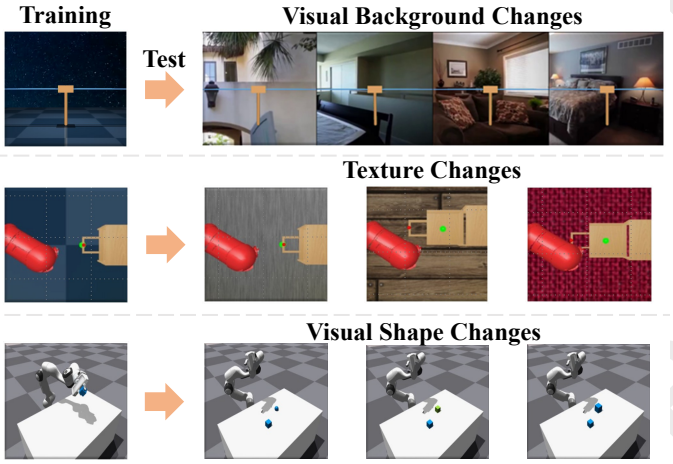


Figure 6: The visualization of DMC-GB, DrawerWorld, and PixMC benchmark.

DrawerWorld [Wang *et al.*, 2021]. We measure generalization on surfaces of different textures, which are unlike the grid texture used for training. These tasks are extremely challenging for two main reasons: 1) The agent has never encountered any realistic textures during training; 2) Each texture has a different color, requiring the agent to handle both color changes and texture changes simultaneously.

PixMC [Xiao *et al.*, 2022]. We also evaluate our method on the PixMC benchmark, which is different from DMC-GB and MetaWorld. These tasks vary in interaction type and difficulty, and there is variability in objects and locations between different episodes.

5.2 Baselines

We compare DisPIM with three different types of baseline methods: the LFS methods, the ensemble methods, and the LFP methods. The LFS methods include: (1) **SAC** [Haarnoja *et al.*, 2018]: a widely used off-policy RL algorithm; (2) **DrQ-v2** [Yarats *et al.*, 2021]: the state-of-the-art LFS algorithm in terms of continuous control; The Ensemble methods include: (3) **SpawnNet** [Lin *et al.*, 2023]: the state-of-the-art method in terms of generalization through ensemble learning; The LFP methods include: (4) **PIE-G** [Yuan *et al.*, 2022]: another state-of-the-art method for generalization by using a pretrained ResNet encoder. (5) **FTD** [Chen *et al.*, 2024]: a state-of-the-art generalization reinforcement learning method that employs the SAM model [Kirillov *et al.*, 2023], enabling the agent to make decisions based solely on task-relevant objects.

For a fair comparison, all baselines employ the same data augmentation method (random shift and overlay augmentation). All experimental results are the average of five random seeds.

[2] $\bar{Q}_{std} = \sqrt{\frac{1}{2}\left(\left(Q_{\bar{\theta}_1} - \mu_Q\right)^2 + \left(Q_{\bar{\theta}_2} - \mu_Q\right)^2\right)}$, where $\mu_Q = \frac{Q_{\bar{\theta}_1} + Q_{\bar{\theta}_2}}{2}$

DMC-GB (video hard)	DrQ-v2	SpawnNet	PIE-G	FTD	Ours
Cartpole, Swingup	130±3	352±63	401±21	207±26	**604±59**
Walker, Stand	151±13	885±36	852±56	421±138	**932±46**
Walker, Walk	34±11	623±73	600±28	396±49	**701±31**
Cup, Catch	97±27	799±121	786±47	658±87	**813±24**
Cheetah, Run	23±5	161±27	154±17	229±43	**223±15**
Finger, Spin	21±4	820±18	762±59	591±146	**916±43**
Average	76	607	592	417	**698**

Table 1: Generalization results on the DMC-GB benchmark. Mean and std of 5 seeds are reported. Our method are robust on visual background changes.

Setting	SAC	DrQ-v2	SpawnNet	PIE-G	FTD	Ours
Training	**100%**	98%	98%	99%	98%	99%
Metal	0%	46%	76%	70%	71%	**86%**
Wood	0%	32%	64%	56%	58%	**70%**
Blanket	0%	8%	76%	71%	68%	**78%**
Average	25%	46%	79%	74%	74%	**83%**

Table 2: Generalization on DrawerWorld benchmark. Evaluation on tasks with distracting textures. Our method is robust to texture changes.

5.3 Simulation

Generalization on DMC-GB Benchmark (visual background changes). The generalization capability of DisPIM and the baselines were evaluated on the DMC-GB benchmark. We evaluate DisPIM on the challenging generalization setting: *video hard*, and compare with several recent state-of-the-art baselines. Results are shown in Table 1. We find that DisPIM outperforms state-of-the-art methods in all instances.

LFS methods (DrQ-v2) hardly acquire any improvement during the test phase. This indicates that when facing environments distracted by task-irrelevant backgrounds, the LFS representations are insufficient to assist the model in solving the task. The ensemble method (SpawnNet) simultaneously employs LFS and LFP encoders. Therefore, SpawnNet demonstrates effectiveness in generalization tasks. However, the parameter count and computational load of the ensemble are substantial, making it challenging to implement in practical applications. LFP methods (FTD and PIE-G) also significantly outperform the LFS method in generalization tasks. However, due to their lack of task-specific capability, the potential of these methods is constrained. DisPIM uses only the LFS encoder for generalization. Therefore, DisPIM not only performs well in all tasks but also has fewer parameters than these methods.

Generalization on DrawerWorld Benchmark (visual texture changes). We also tested our DisPIM on the DrawerWorld benchmark with different background textures. The action space contains the end-effector positions in 3D. We use the success rate as the evaluation metric for its goal-conditioned nature. From Table 2, we observe that DisPIM can achieve better or comparable generalization performance in all settings.

Vision encoder sensitivity to textures poses a big challenge for visual adaptation. 1) LFS methods, i.e., SAC and DrQ-v2, break down during texture testing, as textures are not utilized during the training process. 2) LFP (PIE-G and FTD) and ensemble methods (SpawnNet) have learned the textures, shapes, colors, and other features of images during the pretraining process. These methods are able to leverage the generalization capability of the pretrained representations to recognize and understand these new textures. 3) DisPIM and SpawnNet both utilize pretrained and LFS Encoders. However, SpawnNet only combines the pretrained representations and LFS representations in equal proportions, while our method employs Q-dynamic feature distillation enabling dynamic transfer learning. Consequently, DisPIM demonstrates strong generalization capability with low computational cost, as shown in Figure 2.

Setting	SAC	DrQ-v2	SpawnNet	PIE-G	FTD	Ours
Training	43%	74%	**91%**	90%	83%	**91%**
Shape	0%	52%	86%	82%	71%	**87%**
Color	0%	41%	**77%**	74%	68%	75%
Size	0%	28%	61%	59%	44%	**68%**
Average	11%	49%	78%	76%	67%	**81%**

Table 3: Generalization on PixMC benchmark. Evaluation on tasks with a distractor object. Our method is robust against the distractor object.

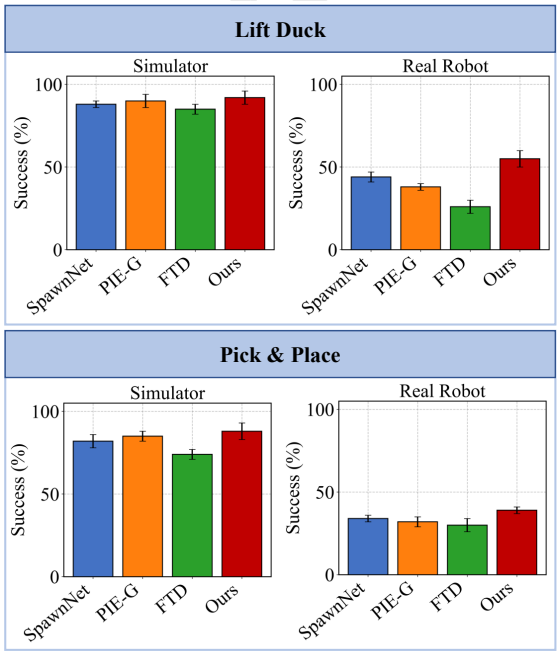


Figure 7: Real world manipulation results. We evaluate each method on each task over 20 trials. We report the mean and standard error.

Generalization on PixMC Benchmark (visual shape changes). We introduce a distractor at test time that varies from the training object in color, shape, and size. Table 3 shows that our DisPIM outperforms baselines in the Franka environment. In particular, DisPIM improves the agent's generalization capability with respect to various colors, shapes,

and sizes, while LFS methods could barely generalize to these changes. We attribute this to the LFS encoder's lack of feature learning capability for shape variations. While ensemble methods and LFP approaches possess the capability to recognize variations in object shapes, their task success rates are inferior to our method, owing to the absence of task-specific knowledge within their encoders.

5.4 Real Robot Experiments

We conduct real robot experiments using an Aubo i5 robot. We perform online training on the simulator and deployed the model directly on the Aubo i5 robot to conduct experiments on the Lift Duck and Pick&Place tasks (shown in Figure 1).

The experimental results are presented in Figure 7. For each baseline, we conduct 20 trials per task and recorded the success rates. We observe that DisPIM performs better than baselines in both the simulator and real robot. On the one hand, for tasks with random locations (Lift Duck), all baselines fail to achieve 50% success rates while DisPIM achieves greater than 50%. Because DisPIM has the benefits from ensemble training with UCB exploration. On the other hand, for tasks with fix location but complex visual background (Pick&Place), all the algorithms fail to achieve 50% success rate. This is because the model has not been fine-tuned on a real robot, direct deployment of the model results in very low success rates for all algorithms.

5.5 Ablation Study

We conduct a series of ablation studies to take a close at the proposed method.

Effectiveness of UCB exploration

To verify the effects of UCB exploration in our framework, we evaluate it on the PixMC manipulation task because this environment is relatively demanding for exploration. We consider a variant of DisPIM, which selects actions without UCB exploration during training. As shown in Figure 8, DisPIM with UCB exploration (green curve) significantly improves the sample efficiency in the environment.

Effect of student model size

We analyze the effect of student model size on the Franka environment from PixMC. Figure 8 shows that the performance of the agent can be improved by increasing the model size, but the improvement is saturated around 6 layers. Thus, we use a 6-layer student encoder for all experiments.

Adopting the teacher encoder for control

We analyze the effect of directly using the teacher encoder for visuo-motor control after training. Figure 8 shows that the teacher encoder achieves a suboptimal control performance in the Franka environment. This result is consistent with the simulation experiment results in the simulation section. Due to its lack of task-specific capability, the potential of the LFP encoder is constrained.

Adopting other PIMs as the teacher model

We investigate the efficacy of other PIMs as a teacher model in our DisPIM. MoCo-v3 [Chen *et al.*, 2021] is a pretrained ViT optimized via contrastive learning. Table 4 shows that DisPIM with MoCo-v3 can also obtain a comparable performance. This result demonstrates the flexibility of the DisPIM framework, which can flexibly apply PIMs as teacher models for feature distillation.

Task	DisPIM	DisPIM (w / MoCo-v3)
Walker Walk	**701** $\pm$ **31**	688 $\pm$ 23
Cheetah Run	**223** $\pm$ **15**	201 $\pm$ 19
Walker Stand	**931** $\pm$ **50**	925 $\pm$ 51

Table 4: Using other PIMs as the teacher model.

Task	ImageNet	CLIP	Ego4D
Walker Walk	701 $\pm$ 31	**720** $\pm$ **30**	451 $\pm$ 19
Cheetah Run	**223** $\pm$ **15**	181 $\pm$ 62	126 $\pm$ 11
Walker Stand	**931** $\pm$ **50**	904 $\pm$ 39	770 $\pm$ 55

Table 5: Using other datasets to pretrain teacher model.

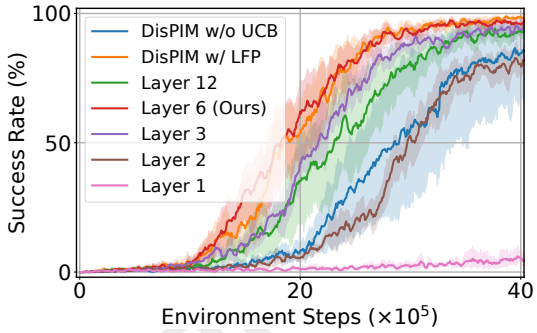


Figure 8: Learning curves of DisPIM with varying values of student encoder (LFS encoder) size on the Franka generalization tasks.

Adopting other datasets for the teacher model

In addition to ImageNet [Deng *et al.*, 2009], we also utilize another widely recognized dataset to pretrain the teacher model. These datasets include CLIP [Radford *et al.*, 2021] (Contrastive Language-Image Pretraining) and Ego4D [Grauman *et al.*, 2021] (Daily-life activity videos). Table 5 shows that the agent pretrained with CLIP achieves comparable performance with those pretrained with ImageNet. Since Ego4D collects the data with the first-person view, the view difference between the tasks and dataset leads to a decrease in performance.

6 Conclusion

We propose DisPIM, a compact and effective framework that enables the control model to leverage both readily available visual priors and task-specific information while minimizing computational costs. Our carefully designed DisPIM framework, combined with UCB exploration, not only enhances performance on training tasks but also crucially preserves generalization capability to unseen environments. DisPIM achieves superior performance over state-of-the-art methods across three challenging benchmarks and real-world robotic experiments.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 62272497 to Hejun Wu).

References

[Banino *et al.*, 2021] Andrea Banino, Adrià Puidomenech Badia, Jacob Walker, Tim Scholtes, Jovana Mitrovic, and Charles Blundell. Coberl: Contrastive bert for reinforcement learning. *arXiv preprint arXiv:2107.05431*, 2021.

[Bellman, 1957] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *International Conference on Computer Vision*, pages 9620–9629, 2021.

[Chen *et al.*, 2022a] Yudong Chen, Sen Wang, Jiajun Liu, Xuwei Xu, Frank de Hoog, and Zi Huang. Improved feature distillation via projector ensemble. *ArXiv*, abs/2210.15274, 2022.

[Chen *et al.*, 2022b] Yudong Chen, Sen Wang, Jiajun Liu, Xuwei Xu, Frank de Hoog, and Zi Huang. Improved feature distillation via projector ensemble. *ArXiv*, abs/2210.15274, 2022.

[Chen *et al.*, 2024] Chao Chen, Jiacheng Xu, Weijian Liao, Hao Ding, Zongzhang Zhang, Yang Yu, and Rui Zhao. Focus-then-decide: Segmentation-assisted reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2024.

[Dai *et al.*, 2022] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. In *ArXiv*, 2022.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Du *et al.*, 2023] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. *arXiv preprint arXiv:2302.06692*, 2023.

[Ebert *et al.*, 2021] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[Grauman *et al.*, 2021] Kristen Grauman, Andrew Westbury, and et al. Ego4d: Around the world in 3,000 hours of egocentric video. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18973–18990, 2021.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, P. Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *ArXiv*, abs/1801.01290, 2018.

[Hansen and Wang, 2020] Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. *2021 IEEE International Conference on Robotics and Automation*, pages 13611–13617, 2020.

[He *et al.*, 2015] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2015.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[Hinton *et al.*, 2015] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[Huang *et al.*, 2024] Xiaohui Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. *ArXiv*, abs/2402.03241, 2024.

[Khandelwal *et al.*, 2022] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *International Conference on Computer Vision*, pages 3992–4003, 2023.

[Lin *et al.*, 2023] Xingyu Lin, John So, Sashwat Mahalingam, Fangchen Liu, and P. Abbeel. Spawnnet: Learning generalizable visuomotor skills from pre-trained networks. *ArXiv*, abs/2307.03567, 2023.

[Luo *et al.*, 2021] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *Neurocomputing*, 508:293–304, 2021.

[Mehta and Rastegari, 2022] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *ArXiv*, abs/2206.02680, 2022.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *ArXiv*, abs/1312.5602, 2013.

[Osband *et al.*, 2016] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Neural Information Processing Systems*, 2016.

[Pei *et al.*, 2023] Renjing Pei, Jian zhuo Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[Romero *et al.*, 2014] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2014.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

[Seo *et al.*, 2022] Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, pages 19561–19579, 2022.

[Shah and Kumar, 2021] Rutav Shah and Vikash Kumar. Rrl: Resnet as representation for reinforcement learning. *ArXiv*, abs/2107.03380, 2021.

[Shridhar *et al.*, 2022] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[Tassa *et al.*, 2018] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *ArXiv*, abs/1801.00690, 2018.

[Tung and Mori, 2019] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. *International Conference on Computer Vision*, pages 1365–1374, 2019.

[Wang *et al.*, 2021] Xudong Wang, Long Lian, and Stella X. Yu. Unsupervised visual attention and invariance for reinforcement learning. *IEEEConference on Computer Vision and Pattern Recognition*, pages 6673–6683, 2021.

[Wang *et al.*, 2022] Che Wang, Xufang Luo, Keith W. Ross, and Dongsheng Li. Vrl3: A data-driven framework for visual deep reinforcement learning. *ArXiv*, abs/2202.10324, 2022.

[Wang *et al.*, 2023] Ziyu Wang, Yanjie Ze, Yifei Sun, Zhecheng Yuan, and Huazhe Xu. Generalizable visual reinforcement learning with segment anything model. *ArXiv*, abs/2312.17116, 2023.

[Wu *et al.*, 2022] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. *ArXiv*, abs/2207.10666, 2022.

[Xiao *et al.*, 2022] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[Yang *et al.*, 2023] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *ArXiv*, abs/2302.03024, 2023.

[Yarats *et al.*, 2021] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *ArXiv*, abs/2107.09645, 2021.

[Yen-Chen *et al.*, 2020] Lin Yen-Chen, Andy Zeng, Shuran Song, Phillip Isola, and Tsung-Yi Lin. Learning to see before learning to act: Visual pre-training for manipulation. In *IEEE International Conference on Robotics and Automation*, pages 7286–7293. IEEE, 2020.

[Yuan *et al.*, 2022] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022.