WHAT MAKES PRE-TRAINED VISUAL REPRESENTATIONS SUCCESSFUL FOR ROBUST MANIPULATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Inspired by the success of transfer learning in computer vision, roboticists have investigated visual pre-training as a means to improve the learning efficiency and generalization ability of policies learned from pixels. To that end, past work has favored large object interaction datasets, such as first-person videos of humans completing diverse tasks, in pursuit of manipulation-relevant features. Although this approach improves the efficiency of policy learning, it remains unclear how reliable these representations are in the presence of distribution shifts that arise commonly in robotic applications. Surprisingly, we find that visual representations designed for manipulation and control tasks do not necessarily generalize under subtle changes in lighting and scene texture or the introduction of distractor objects. To understand what properties *do* lead to robust representations, we compare the performance of 15 pre-trained vision models under different visual appearances. We find that emergent segmentation ability is a strong predictor of out-of-distribution generalization among ViT models. The rank order induced by this metric is more predictive than metrics that have previously guided generalization research within computer vision and machine learning, such as downstream ImageNet accuracy, in-domain accuracy, or shape-bias as evaluated by cue-conflict performance. We test this finding extensively on a suite of distribution shifts in ten tasks across two simulated manipulation environments. On the ALOHA setup, segmentation score predicts real-world performance after offline training with 50 demonstrations.

1 INTRODUCTION

In spite of vast progress in computer vision, the question of how to learn a good visual representation for robotics remains open (Chen* et al., 2021). Elsewhere in computer vision, internet datasets are retrofit to new tasks with transfer learning, which promises both generalization and fast adaptation to downstream tasks in exchange for large-scale pre-training. But in the field of robotics, this promise has yet to be fulfilled even though policies learned from pixels struggle substantially with data efficiency (Cobbe et al., 2018) and especially generalization under visual changes in a scene (Cobbe et al., 2019a).

Recent work (Damen et al., 2018; Grauman et al., 2022) posits that the missing piece is a large pre-training dataset of object interactions across diverse environments — the ImageNet (Deng et al., 2009) or CommonCrawl (Raffel et al., 2020) of manipulation. That is, if we want to improve the visual generalization ability of pre-trained models we simply need to collect datasets of this kind at scale. Indeed, training on large datasets of first-person human interaction data increases policy performance and learning efficiency downstream (Nair et al., 2022; Xiao et al., 2022), but these evaluations occur in environments that are very similar to those used for policy learning. Robotic applications commonly contain environments with varying lighting conditions, scene textures, and background objects, and we want pre-trained representations to allow the robot to handle such variability. Yet we have few concrete measures of how well pre-trained representations generalize out-of-distribution. To take a step towards understanding these problems, our goal in this paper is to thoroughly answer the questions *"which models generalize?"* and *"how can we predict how well a pre-trained model will generalize?"*

Our first key finding is that, when evaluated under visual distribution shifts, models that are designed for manipulation and control do not outperform standard visual pre-training methods. This finding violates our intuitions about what is needed to scale up robot learning and brings into question what constitutes relevant data, how to quantify useful features, and the importance of design choices such as model architecture. In other words, we need more guiding principles to help us understand what representations are good for manipulation and make the problem of iterating on pre-training strategies much more straightforward. Currently, evaluating a pre-trained policy requires training and rolling out downstream policies across multiple environments and experimental conditions. Instead, we can take inspiration from computer vision, which has developed proxies for robust performance on vast out-of-distribution datasets (Geirhos et al., 2021).

Our second key finding is that the emergent segmentation ability of a ViT model is a strong predictor of out-of-distribution generalization performance. We visualize this phenomenon, which we refer to as "spatial features," in Figure 1. Other metrics of model quality, such as linear probes on ImageNet (Chen et al., 2020), and metrics of out-of-distribution performance, such as in-domain accuracy (Miller et al., 2021) and shape-bias (Geirhos et al., 2019), are not predictive for this model class, despite their predictive power in other commonly-studied domains like image classification. This hints at the possibility that the transfer setting of manipulation differs from computer vision tasks typically studied within the robustness literature.

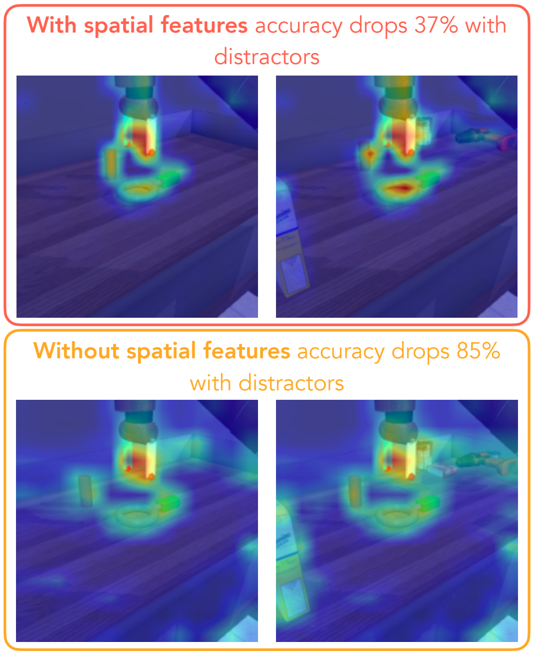
With spatial features accuracy drops 37% with distractors
Without spatial features accuracy drops 85% with distractors

Figure 1: We find that the emergent segmentation ability of ViT attention heads (measured by Jaccard index) predicts performance under visual distribution shift. We refer to models with this property as having "spatial features." Notice how the attention of MVP shifts towards the sugar box distractor object in the bottom right image. The attention of DINO on the top shifts less. The impact of this factor overshadows other design choices such as data relevance.

To reach the conclusions above, we run 9,000 different simulated evaluations. Our simulated environments are adapted from two different existing visual distribution shift benchmarks (Xing et al., 2021; Xie* et al., 2023) to capture the shifts that arise commonly in robotics applications: changes in lighting, background and object texture, and the appearance of distractors. More specifically, we train policies on top of 15 pre-trained models, including 4 models designed for manipulation or control: R3M (Nair et al., 2022), two MVP variants (Xiao et al., 2022; Radosavovic et al., 2022), and VIP (Ma et al., 2022). We further validate these findings by comparing a model designed for manipulation against a model with a similar parameter count on a real-world screwdriver pick-up task using the ACT training framework (Zhao et al., 2023). Through these experiments, we make two striking findings: (1) pre-trained visual models designed for control do not necessarily generalize better than models pre-trained on more standard computer vision datasets and (2) the emergent segmentation performance of a ViT model is a strong predictor of the out-of-distribution generalization of a down-stream policy.

2 Related Work

Representation learning for manipulation. The correct approach to visual representation learning for robotics is still an open question. There is evidence that separating visual representation learning from policy learning can further improve performance (Pari et al., 2022; Parisi et al., 2022). Recent works have shown that models pre-trained on large manipulation-relevant datasets (Goyal et al., 2017; Damen et al., 2018; Shan et al., 2020; Grauman et al., 2022) or learned with visual affordances

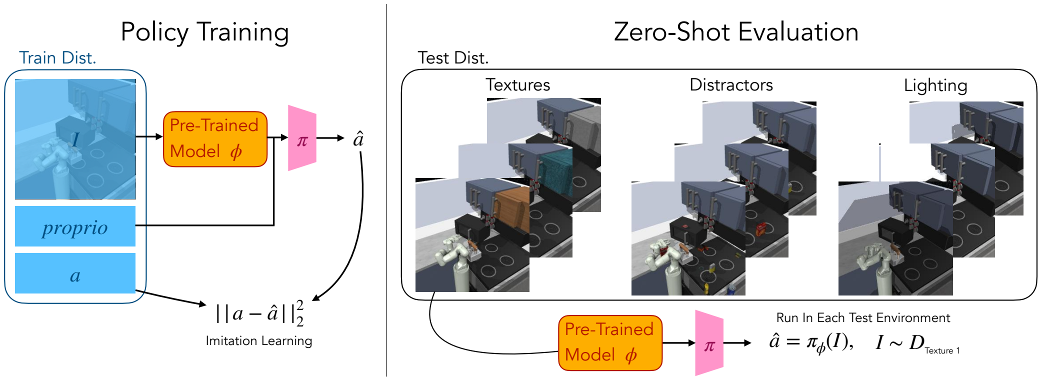

Figure 2: **Evaluation Scheme.** We begin our evaluation procedure by training a policy with behavior cloning on top of frozen features. In every experimental setting, we ablate the encoder used to extract features from the image observation. The learned policy is then evaluated in each of the visual shift environments to attain a zero-shot success value.

from RGBD data (Yen-Chen et al., 2020) can improve the efficiency and performance of policy learning (Karamcheti et al., 2023) in comparison to standard vision datasets such as ImageNet (Deng et al., 2009), but they do not focus on performance under visual distribution shift. We evaluate the performance of R3M (Nair et al., 2022), MVP (Xiao et al., 2022; Radosavovic et al., 2022), and VIP (Ma et al., 2022). Other work has studied generalization of pre-trained representations to new reinforcement learning tasks for manipulation (Ma et al., 2022) and navigation (Sax et al., 2018) where the agent is able to train on visual data from the new environment. Separate from the question of pre-training visual representations is the question of how to best train policies on top of pixel observations (Laskin et al., 2020b; Yarats et al., 2021). Majumdar et al. (2023) benchmarks the performance of pre-trained visual representations on a handful of manipulation environments, but they focus on in-domain performance and also investigate navigation environments. Hu et al. (2023) shows that model performance is highly sensitive to downstream policy learning strategy. We use imitation learning for our evaluation protocol, which they find to be a more stable measure of performance.

Robustness in computer vision. There is extensive work studying the impact of design choices, such as architecture, loss, and data, on the performance of visual models under distribution shift. See Geirhos et al. (2021) for a comprehensive comparison. Most relevant to our paper are studies of shape-bias and architecture. While shape-biased models tend to be more robust than texture-biased ones (Geirhos et al., 2019), the impact of architecture on robustness is less straightforward. For example, vision transformers exhibit better robustness to universal adversarial attacks (Shao et al., 2022), but they are more susceptible to patch-level attacks (Fu et al., 2022). When compared on natural distribution shifts (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a;b), vision transformers and convolutional networks achieve comparable performance when provided with enough data (Bhojanapalli et al., 2021). But for occlusions specifically, vision transformers appear to have an edge (Naseer et al., 2021). Miller et al. (2021) studies the predictive power of in-domain performance for out-of-distribution generalization. Unlike all of these prior works, we focus on how pre-trained representations affect robustness in downstream robotics tasks, instead of downstream vision tasks.

Learning robust policies. Unlike work that focuses on changes in dynamics or initial state distribution (Huang et al., 2021; Raileanu et al., 2020; Laskin et al., 2020a; Cobbe et al., 2019b; Packer et al., 2018; Farebrother et al., 2018), we focus exclusively on the setting of visual distribution shifts. Kirk et al. (2021) and Zhao et al. (2019) provide a comprehensive survey on non-visual distribution shifts in decision making problems. Policy adaptation approaches enable visual robustness specifically by leveraging insights from domain adaptation during policy training (Hansen & Wang, 2021; Fan et al., 2021; Yoneda et al., 2021) or during deployment (Hansen et al., 2021). Other policy adaptation approaches blend pre-training together with reinforcement learning across diverse visual environments (Yuan et al., 2022). In the special case of closing the sim-to-real domain gap, a popular approach is to add randomized textures while training in simulation (Sadeghi & Levine, 2017; Tobin

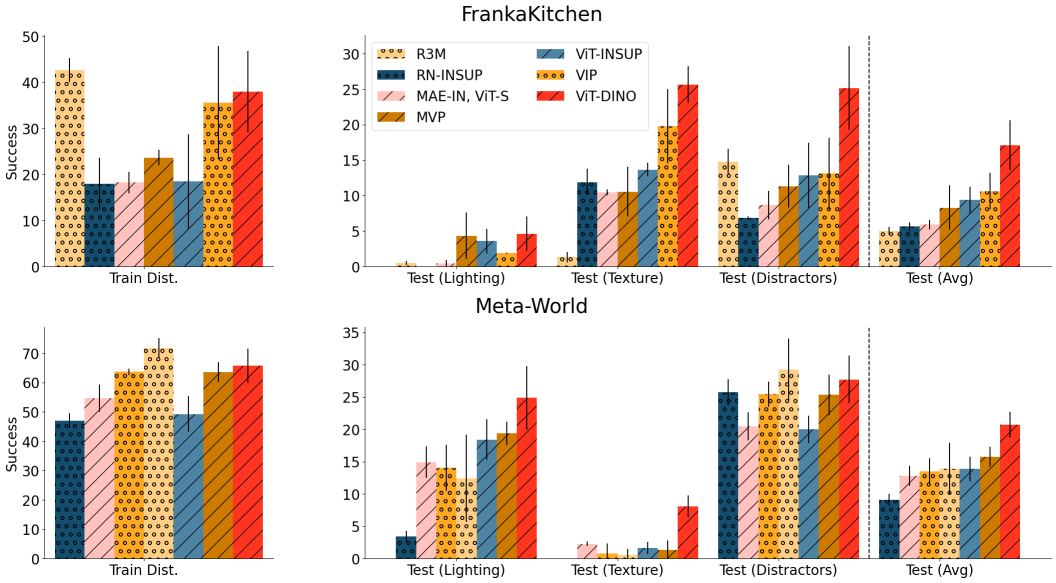

Figure 3: **Visual Generalization Performance.** Models trained with supervision on ImageNet are shades of blue. Models trained with self-supervision on ImageNet are in red. Models trained explicitly for manipulation and control tasks are orange. Dotted bars denote ResNets and slashed bars denote ViTs. Surprisingly, the best performing models are not necessarily the ones designed for manipulation. Each bar is an average over 30 experimental conditions.

et al., 2017; Peng et al., 2018; James et al., 2019). By contrast, our work is interested in explaining properties of a robust visual model for control. Consequently, our insights can be leveraged with or without any task specific data.

3 Environments, Evaluation Protocol, and Pre-Trained Models

Our goal is to understand how robust existing representations for manipulation are to visual distribution shifts that are realistic in robotic applications. To that end, we learn policies on top of frozen, pre-trained encoders and then evaluate these policies zero-shot under changes in lighting, object and scene texture, and the presence of distractors. These shifts are visualized in Appendix Figure 8 and a high level summary of our evaluation procedure is visualized in Figure 2. In this section, we describe the specifics of the manipulation environments, distribution shifts, and policy training setups.

Environments and tasks. We study ten tasks across two simulated manipulation environments, which are selected based on their popularity in studying learning-based approaches to manipulation. Within FrankaKitchen (Gupta et al., 2020) we evaluate performance on opening a microwave, sliding a cabinet door open, pulling a cabinet open, turning a knob, and turning on a light. Within Meta-World (Yu et al., 2019) we study assembling a ring onto a peg, placing an object between two bins, pushing a button, opening a drawer, and hammering a nail.

Distribution shifts. We construct environments to study out-of-distribution generalization within FrankaKitchen and Meta-World. Within FrankaKitchen, we reimplement the texture and lighting changes from KitchenShift (Xing et al., 2021). Within Meta-World we use texture changes from Xie* et al. (2023) and reimplement the same lighting changes as in FrankaKitchen. In both environments we include three levels of distractors: one, three, and nine YCB objects (Calli et al., 2015). We show average performance on each of these distributions shifts as well as performance on the original training distribution, which samples initial positions of the table and kitchen at random. More details about the implementation and parameterization of the distribution shifts are provided in Section A.3.

Policy training. Policy training is done in the same manner as R3M (Nair et al., 2022). A summary of the evaluation scheme is provided in Figure 2. We train an MLP on top of the pre-trained embedding with imitation learning (IL), which, given actions sampled from expert trajectories, $a \sim \mathcal{D}_{train}$,

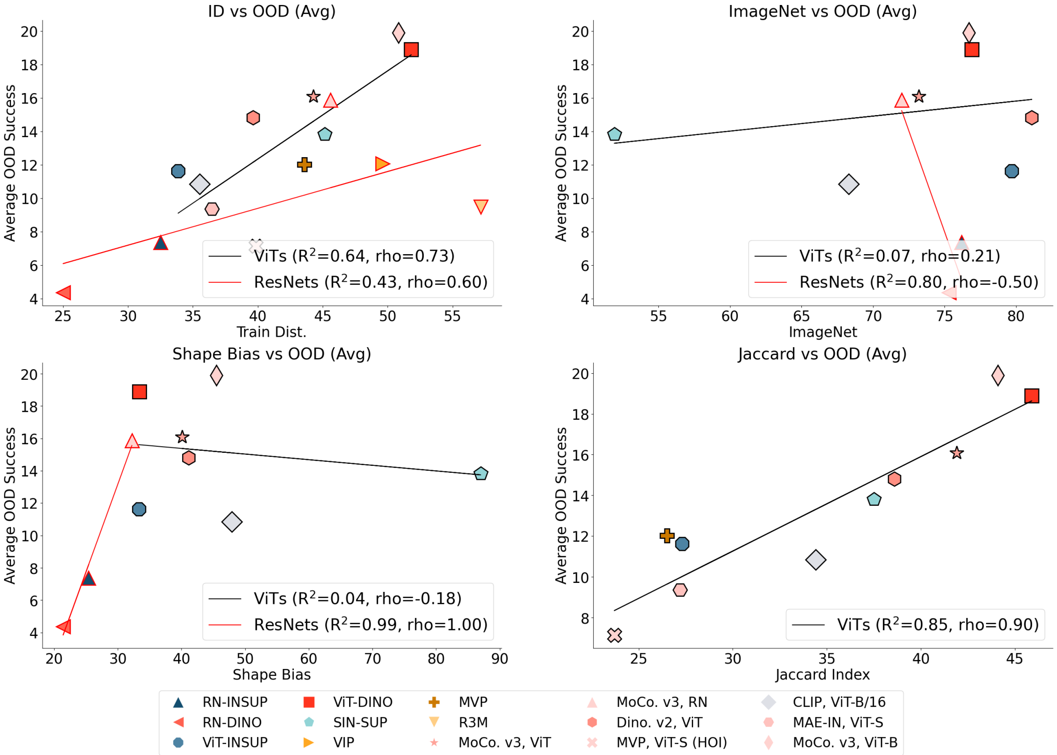

Figure 5: We plot the relationship between different metrics and out-of-distribution (OOD) generalization. There is a promising correlation between shape-bias and OOD performance for ResNets, but not ViTs. Instead, OOD performance for ViTs is strongly correlated with Jaccard index.

minimizes the mean squared error objective, $||a - \hat{a}||_2^2$. Here $\hat{a}$ denotes the action predicted from a given policy. Details of the training procedure are provided in Section A.4. The embedding weights are frozen during policy learning, so the pre-trained models receive no task data. We train 3 different seeds within each task for each of two different camera angles. In total, we learn 60 policies for each model and perform 11 evaluations per policy, including on the train distribution.

Formally, for a pre-trained representation ϕ we learn policies, π_ϕ, each trained with a different seed, camera angle, and task. We average the performance of π_ϕ along each experimental condition and compute the mean performance and error across seeds.

Models. We categorize models by loss type and data source: supervised ImageNet models, self-supervised ImageNet models, and models trained for manipulation and control tasks.

4 Generalization of Models Pre-Trained for Manipulation

One factor motivating work in learning-based robotics is the hypothesis of scale: if we collect more high-quality manipulation data, we should see improvements in policy generalization. However, our understanding of what high-quality data looks like for manipulation and control tasks is still imprecise. Past work on pre-training visual representations for manipulation and control tasks has focused on collecting large object interaction datasets and developing manipulation-relevant losses. But the generalization ability of such models in comparison to standard pre-training methods is still unknown. The goal of this section is to ask: *which models generalize?*

To focus our analysis, we compare models pre-trained for manipulation to two self-supervised ImageNet models and two supervised ImageNet models. Our main result is presented in Figure 3 where we plot the average success rate of the learned policies in the training environment distribution, within each class of visual shift, and across all types of visual shifts. All of the model names as well as the datasets, dataset sizes, model sizes, and loss functions are listed in Appendix Table 2.

We recommend that readers visit this table to get a high level view of each model in our comparison suite.

Models pre-trained for manipulation. Past work has trained visual representations for manipulation in two ways: by training with manipulation-specific losses or on data of human-object interactions. We focus on three recently introduced pre-trained models for manipulation that use different combinations of these approaches: Masked Visual Pretraining (MVP) (Xiao et al., 2022), Reusable Representations for Robot Manipulation (R3M) (Nair et al., 2022), and Value-Implicit Pre-Training (VIP) (Ma et al., 2022). We include important characteristics of these models, including dataset sizes, architecture sizes, and augmentations in Section A.1 and Table 2.

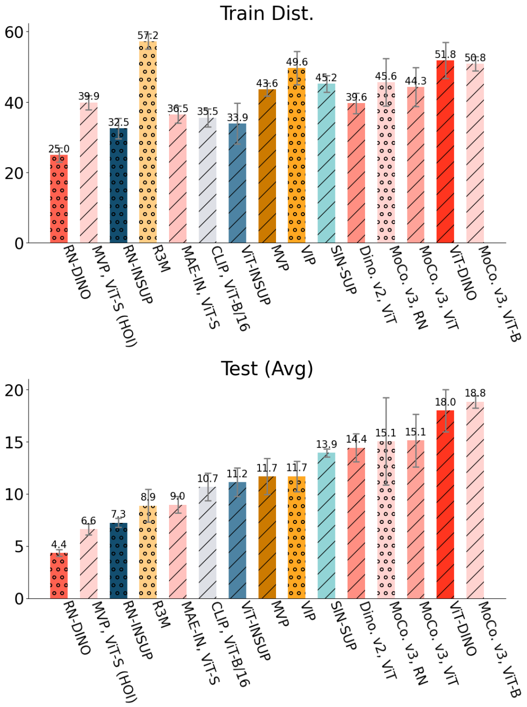

Figure 4: Average success rates for training and test distribution across both environments for every model in our evaluation suite. The best-performing model that was designed for manipulation ranks seventh out of all models evaluated.

These models perform strongly within the training distribution: R3M and VIP in particular comfortably beat standard pre-training baselines. This is expected, especially for R3M which was evaluated on the same training environment. However, under subtle distribution shifts, models designed for manipulation struggle to generalize as well as supervised or self-supervised training with ImageNet. This is surprising for a few reasons. First, each manipulation model is trained on a larger dataset than the pre-trained baselines. Ego4D alone is 4.5M frames while ImageNet is only 1.2M. By parameter count, MVP is also larger than the ViT-S baselines. Finally, we expect human-object interaction datasets such as Ego4D to be more similar to the distribution of images observed when training a manipulation policy. The viewpoints are more varied and the scenes are less curated than ImageNet. Although we expect this to improve the generalization of the learned policy, these results show that other factors may supersede the impact of data relevance or scale alone.

Supervised ImageNet models. Supervised training on ImageNet has long been a baseline for visual pre-training. Past work has found that features learned with supervised learning on ImageNet are also a strong baseline for control: even frozen features are competitive with ground-truth state information on a variety of simulated control tasks (Parisi et al., 2022). However, Parisi et al. (2022) also find that self-supervised learning outperforms supervised learning. Our results contradict this finding. Figure 4 shows that supervised training on Stylized ImageNet achieves a higher success rate in the training distribution than self-supervised training on ImageNet with a masked auto-encoding loss. These models maintain the same rank out-of-domain as well. Even without stylization, in-domain performance of supervised ImageNet models are competitive with models trained with MAE on FrankaKitchen. From these results, we conclude that the presence of supervision is not as predictive of in-domain or out-of-domain performance as other factors. We also find that supervised ImageNet training is still a strong baseline for model generalization: in both settings ViT-INSUP outperforms R3M and MVP.

Self-Supervised ImageNet Models. In Figure 3 we include two self-supervised ViT-S models. Under visual distribution shifts, the model trained with the DINO objective outperforms all three models that are designed for manipulation. Moreover, this trend holds for every distribution shift except Meta-World with distractors. The distractors evaluation suite averages over different levels of distractions and therefore favors models with a high performance in training. In Appendix Section A.8 we plot model performance across different levels of distractors and find that several self-supervised

ViTs experience a smaller drop in performance as more distractors are added compared to ResNet based pre-trained manipulation models like R3M and VIP.

Training with masked autoencoding performs well under distribution shifts in Meta-World, but is less strong under distribution shifts within FrankaKitchen. In Figure 4, we see that MoCo. v3, ViT-B also performs strongly out-of-distribution. When we compare MoCo and DINO against MAE-style training we see that MoCo and DINO use a more extensive set of augmentations. Taking this into account alongside the observation that a ViT trained with supervision on Stylized ImageNet performs well out-of-distribution we conclude that choice of augmentations outweighs the importance of supervision. This extends the findings of Geirhos et al. (2021) to the setting of robust manipulation.

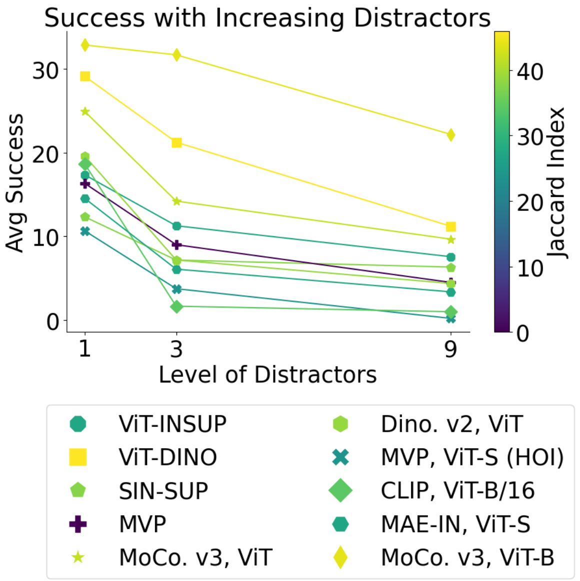

Figure 6: What happens to models with a high Jaccard index under an object-level distribution shift? Surprisingly, the models with the highest Jaccard index maintain the highest performance as the number of distractors increases.

ViTs vs ResNets. One important design choice when selecting a pre-trained model is the choice of architecture. We focus on ResNets and ViTs. In all of our experiments, we use ResNet-50 (He et al., 2016) to be consistent with past work on visual pre-training (Parisi et al., 2022; Nair et al., 2022; Ma et al., 2022). Vision transformers (ViTs) (Dosovitskiy et al., 2021) have seen widespread adoption within computer vision (Khan et al., 2022), but have only recently been used for learning representations for control (Xiao et al., 2022). We find that, on average, ViTs have a slight edge on out-of-distribution generalization compared to equivalently trained ResNets. In Figure 4, out of the seven pre-trained models that perform best out-of-distribution six are ViTs. Ablating architecture alone while holding dataset, training augmentations, and parameter count constant, we can compare the model pairs “MoCo. v3, RN” and “MoCo. v3, ViT”, “RN-DINO” and “ViT-DINO”, and “RN-INSUP” and “ViT-INSUP.” In the latter two pairs, the ViT variant is much stronger out-of-distribution than the ResNet variant. For MoCo, the two variants achieve similar performance out-of-distribution.

Summary. This section identified which pre-trained models generalize, with several interesting findings. First, models designed for manipulaiton do not necessarily perform well under subtle distribution shifts in comparison to more standard pre-training methods. Second, the presence or absence of supervision does not matter as much as other factors on both in- and out-of-distribution generalization. Finally, ViTs have a slight edge over ResNets in out-of-distribution generalization.

5 Properties of Robust Visual Representations for Manipulation

Our findings in the last section are both surprising and somewhat unsatisfying because they contradict many of our intuitions about scale and generalization. In our evaluation suite, we saw that better generalization is not cleanly explained by more data, bigger models, or more relevant data. The goal of this section is to identify the properties of pre-trained models that are predictive of generalization. To that end, we correlate out-of-distribution performance with three metrics that have been previously connected to generalization in the machine learning and computer vision literature—in-domain performance, accuracy of a linear probe trained on ImageNet, and shape-bias. We also include a fourth metric, which is specific to ViTs: the emergent segmentation accuracy of the output attention heads. We describe each metric in detail in Section 5.1, discuss our setup for correlating performance in Section 5.2, and analyze our results in Section 5.3.

5.1 Metrics

ID vs OOD. One of the goals of this paper is to understand how well the findings from existing evaluations of pre-trained models hold under the inevitable environment changes that we expect to see in a real-world setting. If in-distribution performance is reasonably predictive of generalization to our suite of distribution shifts, it is sufficient for researchers to continue developing pre-trained models with existing methods of evaluation. Past work has also shown that the in-distribution performance of a pre-trained model is positively correlated with out-of-distribution performance for a variety of computer vision tasks (Miller et al., 2021). Concretely, we measure in-distribution performance as the success rate of the policy within the training distribution.

Imagenet vs OOD. Training linear probes on Imagenet is a common protocol for evaluating the quality of learned representations (He et al., 2019; Chen et al., 2020). Hu et al. (2023) make the related finding that the ImageNet k-NN accuracy of a pre-trained model is predictive of performance on imitation learning with a visual reward function. We evaluate ImageNet validation set accuracy for all models with linear probes available.

Shape-Bias vs OOD. Shape bias is the extent to which a model makes prediction decisions based on shape. We calculate shape bias as the percent of shape classification decisions out of the set of texture or shape classifications on the Stylized-ImageNet validation set (Geirhos et al., 2019) using the same probes described above.

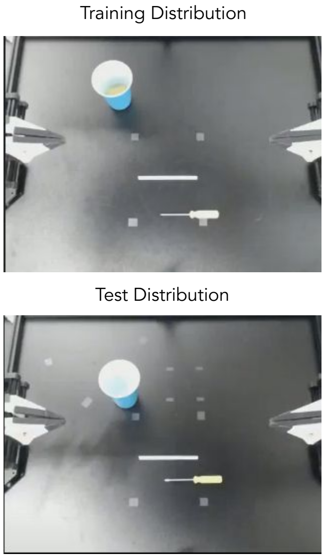

Figure 7: Real world training and test distribution. The test distribution differs from the training distribution in the position of the target objects and the direction of the lighting.

Jaccard vs OOD. Finally, for all of the ViT models, we look at the emergent segmentation performance. We denote this nonlinear, deterministic transform as M. Formally, we compute the Jaccard index by calculating the mIoU on the PASCAL VOC validation set, D_{Pascal}:

$$J(x_i, x_j) = \mathbb{E}_{D_{Pascal}}\left[\frac{A \cap B}{A \cup B}\right]$$

Where A is a shorthand for positive classification for the target class by $M(\phi(\cdot))$ and B is a shorthand for positive label for the target class. J is evaluated pixel-wise over image indices x_i and x_j. We evaluate the Jaccard index of an interpolated attention map averaged across heads in the last attention block at the [CLS] token.

5.2 Setup

We measure the coefficient of determination (R^2) and Spearman's rank correlation (ρ) for the correlation between the out-of-distribution success rate and each metric described above. Our goal is to find a metric that will result in high correlation between the metric and the OOD success, i.e. both coefficients being close to 1.0. We fit separate trend lines to ViTs and ResNets. Because of the lack of available probes, we exclude MVP, MVP ViT-S HOI, R3M, VIP, and MAE-IN ViT-S from the shape bias and ImageNet probe correlations. Each point represents one of the 15 pre-trained models we evaluated and represents the average of 6,000 evaluation runs.

5.3 Results

We visualize the correlation between each metric and the average out-of-distribution success rate in Figure 5. Although we see a positive relationship between in- and out-of distribution generalization, there are pre-trained models that notably deviate from this trend. Among ViT models one example is MVP, ViT-S (HOI): the average success rate of this model drops to 6.63 from 39.86. By contrast, we find that ImageNet accuracy of a linear probe poorly predicts generalization performance for ViTs.

We also see little correlation between shape-bias and OOD performance for ViT models, but a promisingly strong correlation on the subset of ResNets evaluted. This is surprising because humans make highly shape-biased decisions and increasing shape-bias increases the robustness of imagenet trained CNNs (Geirhos et al., 2019; 2021). One explanation of this finding is that the ViT architecture obviates the need for shape-biased features. For example, a ResNet-50 trained with the DINO training scheme has a strong shape-bias, but not the equivalent ViT model.

Finally, we visualize the relationship between the Jaccard index and OOD performance on all ViT models in Figure 5. There is a strong positive correlation between Jaccard index and OOD performance both in terms of rank correlation and the coefficient of determination. These results suggest that while shape-bias may not be predictive of the OOD generalization ability of a pre-trained ViT, the segmentation ability is a predictive alternative.

Model	Success
MVP	0%
MoCo-v3	40%

Table 1: Success rates on the task of picking up the screwdriver.

One counter-argument to the use of Jaccard index as a metric for for OOD performance is that it would be less predictive for object-level distribution shift, which would occur any time a large distractor is placed in the background of the image. In Figure 6, we plot the success rates of each ViT model as the number of objects increases and verify that the models with the higher Jaccard index actually maintain the highest performance as the number of distractors increases.

5.4 Validating in the real world

In this section, we validate our finding on a real-world generalization scenario by comparing a ViT-B model designed for control (MVP) against a model not designed for control but with a high emergent segmentation score (MoCo-v3).

Setup. We learn policies for picking up a screwdriver on the ALOHA setup using the ACT training framework (Zhao et al., 2023). The training dataset is comprised of 50 episodes collected by an expert human demonstrator. Images are collected from 4 camera view points (one on each wrist, one top camera, and one front camera). We replace the standard encoder with a ViT-B and change the initialization of the encoder based on the experimental condition (i.e., we select for a different pre-trained model). We follow the standard ACT training paradigm with the hyperparameters listed in Appendix Table 4. From the training data to the test runs there is a distribution shift in both the placement of the target object (the screwdriver) and in the direction of the lighting. This is visualized in Figure 7. We calculate success on screw pick ups averaged over 10 rollouts in the test environment.

Results. We find that MoCo-v3 is stronger on this setting than MVP, even though it is not explicitly designed for manipulation. We find that the MoCo-v3 initialized encoder is able to achieve a success rate of 40% on this task while the MVP initialized encoder is not able to successfully grasp the target object. Qualitatively, the MVP model fails in localizing the object when attempting the grasp, whereas MoCo-v3 model reliably localizes the object, but experiences more failure in finding the right grasp point.

6 Conclusion

Summary. In this paper, we make several surprising findings about the generalization ability of pre-trained visual representations for manipulation tasks. First, we find that, contrary to the current direction in the literature, models pre-trained on manipulation-relevant data do no necessarily generalize better than models trained on standard pre-training datasets (such as ImageNet). Instead, we uncover a recipe for strong generalization: ViT models with a high emergent segmentation accuracy generalize well under visual distribution shifts. Emergent segmentation accuracy is not only a stronger predictor of generalization than many other metrics for robustness, but also requires no additional training to evaluate. This insight can guide the development of pre-trained vision models in future work: preferring architecture development and training algorithms that lead to strong emergent segmentation as opposed to only training on more manipulation-relevant data.

7 Reproducibility

All of our code is open-sourced and all changes to relevant libraries are available in the supplementary materials. We also include all of the XML files that we used to generate our visual shift scenarios. Our appendix includes the exact hyperparameters we used to conduct our simulated policy training and our real-world experiments.

References

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10211–10221, 2021. doi: 10.1109/ICCV48922.2021.01007.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pp. 510–517, 2015. doi: 10.1109/ICAR.2015.7251504.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

Karl Cobbe, Oleg Klimov, Christopher Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. *ArXiv*, abs/1812.02341, 2018.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *ArXiv*, abs/1912.01588, 2019a.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019b.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Animashree Anandkumar. Secant: Self-expert cloning for zero-shot generalization of visual policies. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3088–3099. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fan21c.html.

Jesse Farebrother, Marlos C. Machado, and Michael H. Bowling. Generalization and regularization in dqn. *ArXiv*, abs/1810.00123, 2018.

Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=28ib9tf6zhr.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Neural Information Processing Systems*, 2021.

Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The “something something” video database for learning and evaluating visual common sense. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5843–5851, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Yu Heng Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2022.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *Conference on Robot Learning*, pp. 1025–1037. PMLR, 2020.

Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *International Conference on Robotics and Automation*, 2021.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations*, 2021.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8320–8329, 2021a. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021b.

Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal, 2023.

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *ArXiv*, abs/2107.02729, 2021.

Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12619–12629, 2019.

Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3505244. URL https://doi.org/10.1145/3505244.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktaschel. A survey of generalisation in deep reinforcement learning. *ArXiv*, abs/2111.09794, 2021.

Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *arXiv preprint arXiv:2004.14990*, 2020a.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119*, 2020b. arXiv:2004.04136.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

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

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. *ArXiv*, abs/2107.04649, 2021.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=o2mbl-Hmfgd.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Xiaodong Song. Assessing generalization in deep reinforcement learning. *ArXiv*, abs/1810.12282, 2018.

Jyothish Pari, Nur Muhammad (Mahi) Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. *ArXiv*, abs/2112.01511, 2022.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Kumar Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8, 2018.

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. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in deep reinforcement learning. *ArXiv*, abs/2006.12862, 2020.

Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image. *ArXiv*, abs/1611.04201, 2017.

Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas J. Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. 2018.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018. URL `http://arxiv.org/abs/1704.06888`.

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9866–9875, 2020.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=lE7K4n1Esk`.

Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/touvron21a.html`.

T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020.

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

Annie Xie*, Lisa Lee*, and Chelsea Finn. Benchmarking environment generalization in robotic imitation learning, 2023. URL `https://github.com/RLAgent/factor-envs`.

Eliot Xing, Abhinav Gupta, Sam Powers*, and Victoria Dean*. Kitchenshift: Evaluating zero-shot generalization of imitation-based policy learning under domain shifts. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. URL `https://openreview.net/forum?id=DdglKo8hBq0`.

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

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 (ICRA)*, 2020. URL `https://yenchenlin.me/vision2action/`.

Takuma Yoneda, Ge Yang, Matthew R. Walter, and Bradly Stadie. Invariance through latent alignment, 2021.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019. URL `https://arxiv.org/abs/1910.10897`.

Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. 12 2022. doi: 10.48550/arXiv.2212.08860.

Chenyang Zhao, Olivier Sigaud, Freek Stulp, and Timothy M. Hospedales. Investigating generalisation in continuous deep reinforcement learning. *ArXiv*, abs/1902.07015, 2019.

Tony Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023.

A Appendix

A.1 Pre-Trained Model Details

RN-INSUP (He et al., 2016) is a ResNet model trained on the ImageNet classificaiton task. We use the default weights and model provided by the Pytorch (Paszke et al., 2019) library.

ViT-INSUP is a Vision Transformer (Dosovitskiy et al., 2021) that has been distilled (Touvron et al., 2021) from a larger network that was trained on the ImageNet classification task. In our experiments, we use the model weights and architecture provided in Naseer et al. (2021) with a patch size of 16.

SIN-SUP (Naseer et al., 2021) trains a vision transformer on Stylized Image-Net (SIN) (Geirhos et al., 2019). The SIN dataset was constructed to increase the degree to which a model makes predictions on shape instead of texture. Our model weights come from Naseer et al. (2021) and we use the non-distilled DeiT (Touvron et al., 2021) training variant.

ViT-DINO (Caron et al., 2021) is trained with extensive augmentations and a self-supervised, contrastive loss that together lead to emergent segmentation within the self-attention heads of the ViT model. We use the model and weights provided by Caron et al. (2021). Interestingly, we don't find the DINO objective to lead to a high shape-bias. This suggests that there are other metrics that measure the degree to which a model is object-centric other than shape-bias.

ResNet50-DINO is learned with the same recipe as ViT-DINO. We use the model and weights from Caron et al. (2021).

MoCo. v3, RN (Chen* et al., 2021) leverages a contrastive loss with momentum encoding (He et al., 2019) of positive targets. It is trained with the same recipe as MoCo. v3, ViT-B.

MoCo. v3, ViT-B (Chen* et al., 2021) are trained in a similar manner as the original MoCo (He et al., 2019), but with changes to improve the stability of training, which are specific to the ViT archiecture. We use the checkpoint after 300 epochs.

MoCo. v3, ViT-S (Chen* et al., 2021) is trained in a similar manner as MoCo. v3, ViT-B. Even though the smaller model benefits from a longer training horizon, we use the checkpoint at 300 epochs for consistency.

MAE-IN, ViT-S follows the same training recipe as MVP, but on top of the ImageNet dataset. We use the weights provided by Radosavovic et al. (2022).

R3M (Nair et al., 2022) trains a ResNet model with a combination of manipulation-specific losses–including a time-contrastive loss (Sermanet et al., 2018), video-language alignemnt loss, and L1-regularization–on the Ego4D (Grauman et al., 2022) dataset.

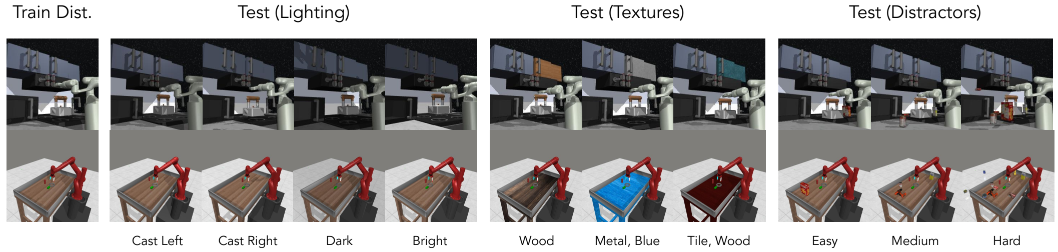

Figure 8: We visualize each distribution shift from the left camera angle on the FrankaKitchen (top) and Meta-World (bottom) environments.

MVP (Radosavovic et al., 2022) trains a ViT-B for masked autoencoding (MAE) (He et al., 2021) on the Ego4D (Grauman et al., 2022), Something-Something (Goyal et al., 2017), YouTube 100 Days of Hands (Shan et al., 2020), EpicKitchens (Damen et al., 2018), and ImageNet (Deng et al., 2009) datasets. Unlike R3M, the model is not designed to be exclusive to manipulation.

MVP, ViT-S (HOI) (Xiao et al., 2022) is a predecessor of the model described above that trains a ViT-S/16 with an MAE objective on Something-Something (Goyal et al., 2017), YouTube 100 Days of Hands (Shan et al., 2020), EpicKitchens (Damen et al., 2018), and ImageNet (Deng et al., 2009).

VIP (Ma et al., 2022) uses an action-free dual of the Algaedice (Nachum et al., 2019) objective to learn representations that are useful for trajectory optimization or reinforcement learning of control tasks unseen during representation pre-training. They train a ResNet-50 on Ego4D with this objective.

CLIP, ViT-B/16 (Radford et al., 2021) uses contrastive language-image pre-training to learn visual representations trained on an extensive internet datsaet. The learned models exhibit strong zero-shot performance for multiple tasks such as image classification.

DiNo v2, ViT (Oquab et al., 2023) scales Caron et al. (2021) to more parameters and a larger dataset. The full model is a 1B parameter ViT trained on LVD-142M, which is a 142M frame dataset composed of ImageNet-1k, ImageNet-22k, Google Landmarks (Weyand et al., 2020), and a collection of other datasets spanning fine-grained classification, segmentation, depth estimation, and retrieval. The full model is distilled into smaller models. We select the ViT-S distilled model for our experiments. In Table 2, we list the augmentations used on the teacher model. The training loop is only lightly modified during distillation. Suprisingly, the v2 model sees worse in- and out-of-domain performance on our evaluation suite in spite of being distilled from a ladrger model trained on a bigger dataset.

A.2 Details of the Environments

FrankaKitchen (Gupta et al., 2019) is a simulated kitchen environment with a 9-DoF Franka robot. There a multiple household objects available for interaction. The environment is designed to compose tasks together hierarchically, but we focus on learning policies to successfully complete a single task. The episode length is 50 and we inherit the randomization scheme used in R3M, which randomizes the position of the kitchen at the start of each episode.

Meta-World (Yu et al., 2019) is a simulated manipulation environment that consists of various table-top manipulation interactions. Unlike FrankaKitchen, the scene objects vary between different tasks. The positions of the objects are randomized at the start of each episode. The maximum episode length is 500.

A.3 Details of the Disribution Shifts

Each distribution shift is visualized from the left camera angle in Figure 8. We don't use the MuJoCo scanned object dataset that is used in (Xie* et al., 2023) because of imperfections in the coloring of the textures.

Name	Loss Function	Architecture	Datasets	Augmentations
RN-INSUP	BCE-Loss	ResNet-50 (23M params)	ImageNet (1.2M frames)	Random crop, Horizontal flip
ViT-INSUP	BCE-Loss	ViT-S/16 (22M params)	ImageNet (1.2M frames)	Random crop, Horizontal flip
SIN-SUP	BCE-Loss	ViT-S/16 (22M params)	Stylized-ImageNet (1.2M frames)	Random crop, Horizontal flip
ResNet50-DINO	Distillation	ResNet-50 (23M params)	ImageNet (1.2M frames)	Multi-crop, Color-jittering, Gaussian blur, Solarization
ViT-DINO	Distillation	ViT-S/16 (22M params)	ImageNet (1.2M frames)	Multi-crop, Color-jittering, Gaussian blur, Solarization
MoCo. v3, RN	Contrastive	ResNet50 (23M params)	ImageNet (1.2M frames)	Resize, Color-jittering, Horizontal flip, Grayscale, Gaussian blur, Solarization
MoCo. v3, ViT-S	Contrastive	ViT-S/16 (22M params)	ImageNet (1.2M frames)	Resize, Color-jittering, Horizontal flip, Grayscale, Gaussian blur, Solarization
MoCo. v3, ViT-B	Contrastive	ViT-B/16 (88M params)	ImageNet (1.2M frames)	Resize, Color-jittering, Horizontal flip, Grayscale, Gaussian blur, Solarization
MAE-IN, ViT-S	Masked auto-encoding	ViT-S (22M params)	ImageNet (1.2M frames)	Random resize, Random crop
R3M	Time-contrastive, L1-regularization, Video-lang alignment	ResNet-50 (23M params)	Ego4D (4.3M frames)	Random crop
MVP, ViT-S (HOI)	Masked auto-encoding	ViT-S (22M params)	EpicKitchens 100 Days of Hands, Something-Something (700k frames)	None
MVP	Masked auto-encoding	ViT-B (88M params)	Ego4D, ImageNet EpicKitchens, 100 Days of Hands, Something-Something (4.5M frames)	None
VIP	Algaedice Dual	ResNet-50 (23M params)	Ego4D (4.3M frames)	Random crop
CLIP, ViT-B/16	Contrastive	ViT-B/16 (88M params)	Internet data (400M pairs)	Random crop
DiNo v2, ViT	Distillation	ViT-S/14 (21M params)	LVD (142M frames)	Multi-crop, Color-jittering, Grayscale, Gaussian blur, Solarization

Table 2: List of pre-trained models with corresponding loss function, augmentations, and datasets used for pre-training. We color code by the data and loss type: ImageNet supervised, self-supervised, trained specifically for manipulation or control tasks, and other.

A.4 Policy Training Details

We learn a 2-layer MLP on top of the pre-trained, frozen features with 10 demonstrations. We use the same expert demonstrations as in R3M. We train policies independently over the 'left_cap2' and 'right_cap2' camera angles and show results averaged over both camera angles. We also provide proprioception to the policy. The final performance is averaged over the task settings for each seed. The hyperparamters for policy training are summarized in Table 3. Error bars are 95% confidence interval over seeds.

Hyperparameter	Value
Loss type	MSE
Learning rate	0.001
Batch size	32
Train steps	20,000
Optimizer	Adam

Table 3: Hyperparameters for IL Policy Training

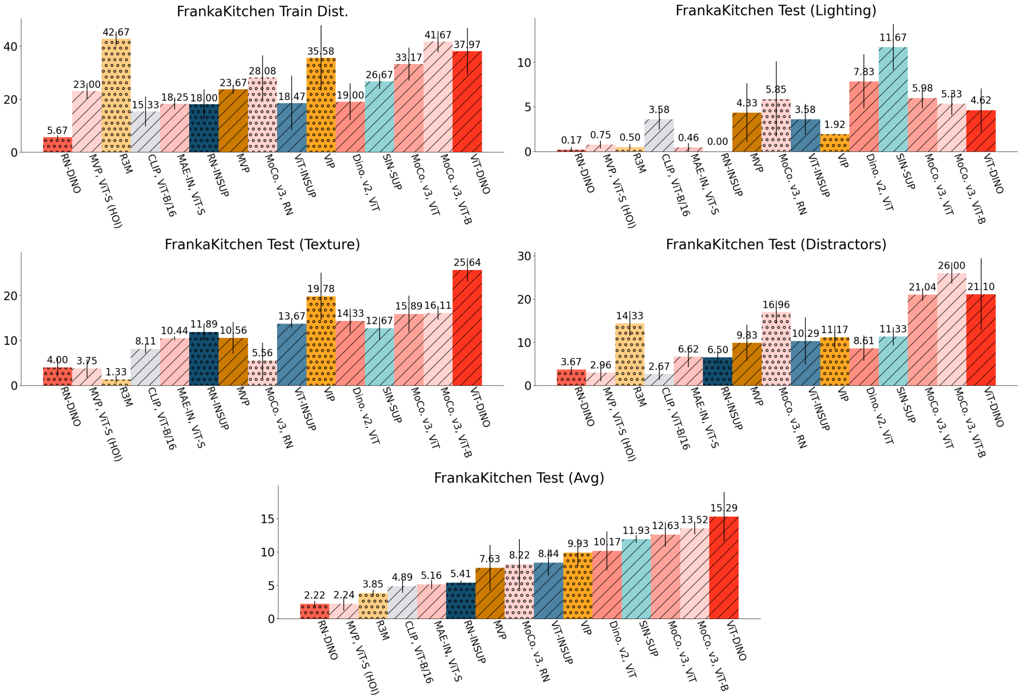

Figure 9: **Detailed OOD Performance on FrankaKitchen.**

A.5 OOD Perf Details

To provide a more granular understanding of how the complete set of models performs on our evaluation suite, we break down performance by distribution shift type and environment in Figures 9 and 10.

A.6 ImageNet vs OOD Details

To evaluate ImageNet accuracy, we use all publicly available probes that have been trained on top of the frozen model features and evaluate them on the ImageNet validation set. The models with available probes are RN-INSUP, RN-DINO, MoCo. v3 RN, ViT-INSUP, ViT-DINO, MoCo. v3 ViT, Dino v2 ViT, MoCo. v3 ViT, SIN-SUP, and CLIP ViT-B/16 and we use the probes that are provided in the implementations cited in Section A.1.

A.7 Shape-Bias Details

We evaluate shape-bias using the 'model-vs-human' evaluation framework from Geirhos et al. (2021) and use the same probes from Section A.6 to get classification results on the cue-conflict validation dataset ($D_{cue-conflict}$). The cue-conflict dataset contains images where the shape and texture cues are in conflict (e.g., a cat with the texture of the elephant). The shape bias of the model is the ratio of classification decisions made based on the shape cue (e.g., cat) vs the texture cue (e.g., elephant).

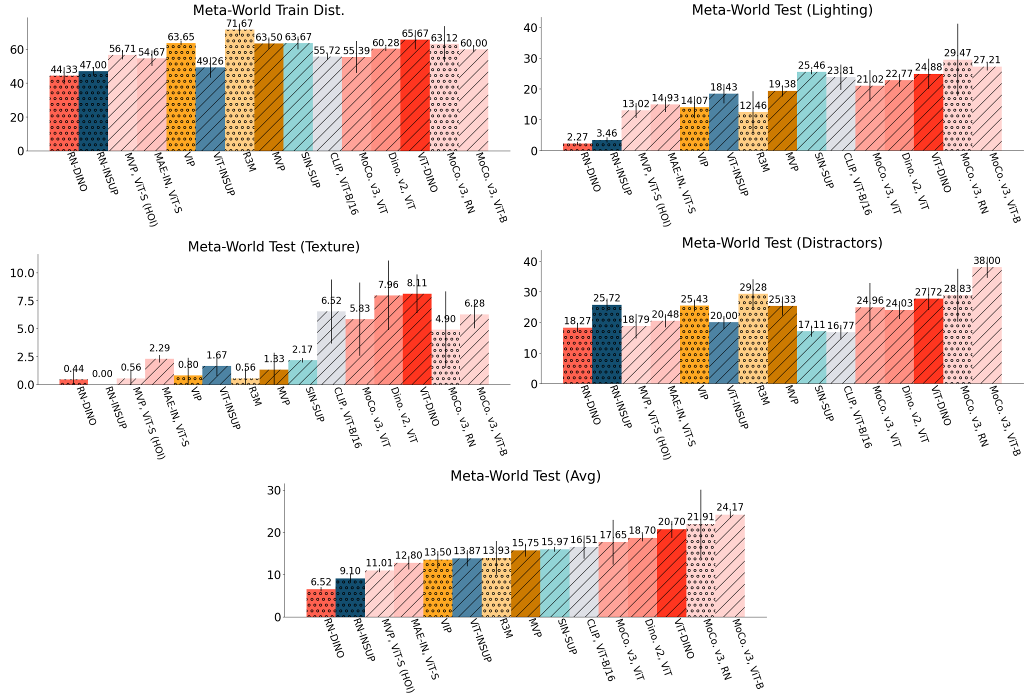

Figure 10: **Detailed OOD Performance on Meta-World.**

Notably, Naseer et al. (2021) find that vision transformers are more shape-biased when making classification decisions than equivalently trained convolutional networks. In our results, we don't find vision transformers to be more strongly shape biased. Vision transformers and convolutional networks vary in how they handle spatial resolution: spatial resolution decreases in each layer of ResNet-50 but remains constant within a ViT. This could explain why we see the ViT architecture somewhat obviating the need for shape-bias in our results.

A.8 Different Levels of Distractors

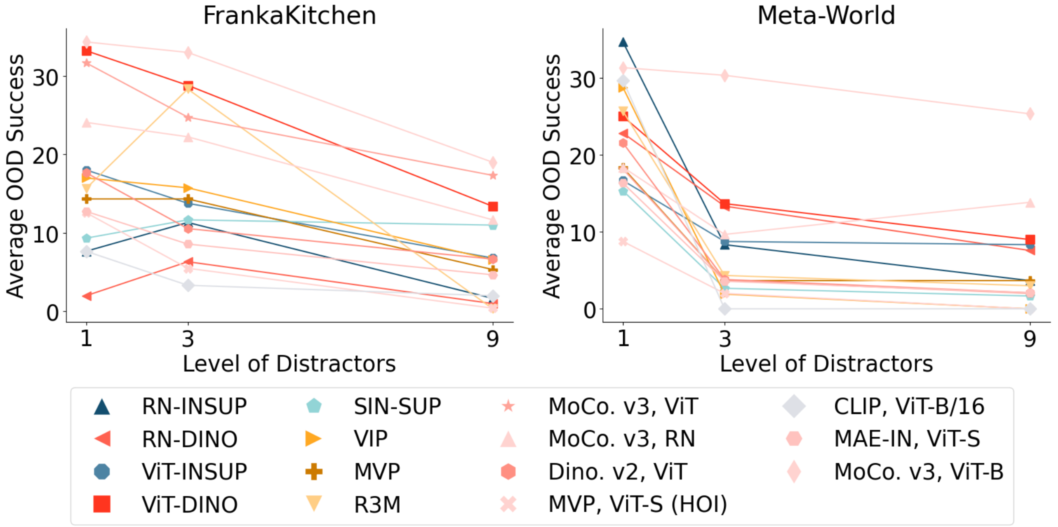

Figure 11: Different levels of distractors.

We extend Figure 6 by including results for ResNets in Figure 11. Models are color coded using the original color scheme in the paper.

A.9 Finetuning

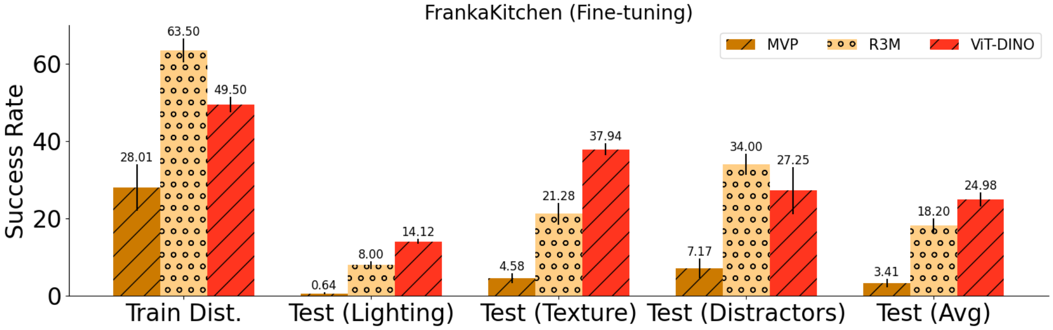

Figure 12: **Finetuning in FrankaKitchen.**

Because the goal of this paper is to probe the quality of learned representations, we follow the tradition of performing evaluation on top of frozen model features. This evaluation is also consistent with the increasing view of pre-trained visual representations as "foundation models" (Bommasani et al., 2022; Oquab et al., 2023) that can be deployed without any gradient updates. Nonetheless, even in the fine-tuning regime, in Figure 12 we still see stronger performance from models that are not designed for manipulation. In this setting, we increased the number of demonstrations to 25 to allow for more data diversity when training the encoders.

A.10 Real-World Experiment Details

Our demonstration data contains two subtasks: an initial screwdriver pick-up and then a handover that happen in sequence. We only evaluate success on the subtask of picking up the screwdriver.

Hyperparameter	Value
Chunk Size	100
KL Weight	10
Batch size	8
Epochs	10,000
Optimizer	Adam
Learning Rate	1e-5

Table 4: Hyperparameters for Policy Training

A.11 Additional Experiments

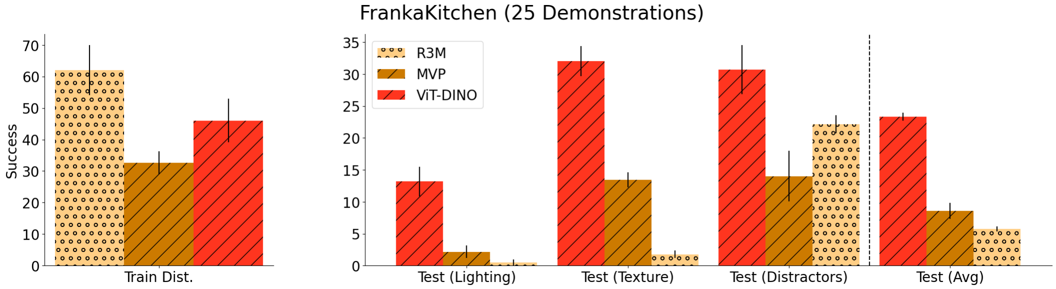

Figure 13: Performance on FrankaKitchen with more demonstrations remains consistent.

Increasing the number of demonstrations. To study the impact of increased demonstrations, we also show the performance of R3M, MVP, and ViT-DINO with 25 demonstrations in Figure 13. This is the largest number of demonstrations used in the R3M evaluation suite. We find that the trend remains consistent, if not exaggerated, with increased demonstrations.

Comparison to PIE-G. PIE-G Yuan et al. (2022) blends pre-training together with reinforcement learning to learn visually robust representations. The pre-trained model used in PIE-G is a ResNet-18 with features extracted from the second layer of the network. We extract these features for comparison in our benchmark and perform average pooling along the spatial dimensions (that is, along the height and width) to produce a 128-dimensional feature. This model achieves a training performance of 0.0 across all the FrankaKitchen training tasks.

Analysing the Jaccard index of GradCAM applied to ResNet models. In our experiments, the Jaccard index was the most predictive metric of out-of-distribution performance. To arrive at an equivalent metric for ResNet models, we evaluate the Jaccard index of segmentation maps generated with Grad-CAM (Selvaraju et al., 2017). To generate our segmentation maps, we use the Grad-CAM implementation made available by Gildenblat & contributors (2021). Figure 14 shows that generating segmentation maps in this way does not give a predictive metric for out-of-distribution performance for ResNets. One explanation for this result is that Grad-CAM is not the best measure of the internal spatial features of a ResNet model. Another hypothesis is that ViTs have the capacity to model shape directly in their attention heads, which obviates the need for shape-biased features. The ResNet model architecture may not have the capacity to support this kind of representaiton, which requires shape bias to be encoded directly in features (Geirhos et al., 2019).

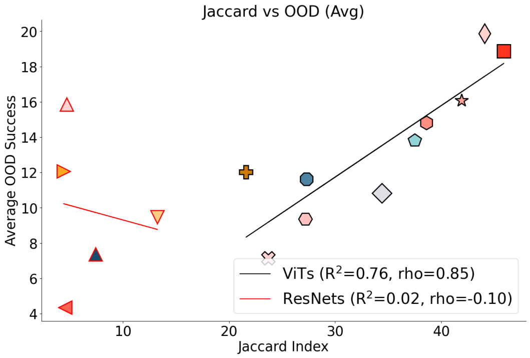

Figure 14: Jaccard index with ResNet models included.