Zero-Shot Visual Generalization in Robot Manipulation

Sumeet Batra
University of Southern California
United States
ssbatra@usc.edu

Gaurav S. Sukhatme
University of Southern California
United States
gaurav@usc.edu

Abstract: Training vision-based manipulation policies that are robust across diverse visual environments remains an important and unresolved challenge in robot learning. Current approaches often sidestep the problem by relying on invariant representations such as point clouds and depth, or by brute-forcing generalization through visual domain randomization and/or large, visually diverse datasets. Disentangled representation learning – especially when combined with principles of associative memory – has recently shown promise in enabling vision-based reinforcement learning policies to be robust to visual distribution shifts. However, these techniques have largely been constrained to simpler benchmarks and toy environments. In this work, we scale disentangled representation learning and associative memory to more visually and dynamically complex manipulation tasks and demonstrate zero-shot adaptability to visual perturbations in both simulation and on real hardware. We further extend this approach to imitation learning, specifically Diffusion Policy, and empirically show significant gains in visual generalization compared to state-of-the-art imitation learning methods. Finally, we introduce a novel technique adapted from the model equivariance literature that transforms any trained neural network policy into one invariant to 2D planar rotations, making our policy not only visually robust but also resilient to certain camera perturbations. We believe that this work marks a significant step towards manipulation policies that are not only adaptable out of the box, but also robust to the complexities and dynamical nature of real-world deployment. Supplementary videos are available at https://sites.google.com/view/vis-gen-robotics/home.

Keywords: manipulation, representation learning, robot learning

1 Introduction

A key requirement of any generalist robot system deployed in the real-world is the ability to perform tasks across visually diverse environments. High-dimensional inputs like RGB images offer rich information but also introduce complexity due to the curse of dimensionality. Given the enormous diversity of real-world visual data, accounting for every possible variation within a fixed dataset is intractable. Extracting the underlying structural knowledge of the world from visual data while being robust to semantically irrelevant visual perturbations remains an open question. The robot learning field has largely relied on one of several trends, one of which is to train agents in simulation, where visual complexity can be controlled and large-scale synthetic and diverse data can be generated efficiently through GPU-accelerated simulators [1, 2, 3]. However, transferring policies trained in simulation to the real world is hindered by the "Sim2Real" gap caused by mismatches in fidelity and unmodeled dynamics. Domain randomization is the leading strategy to close this gap by varying the simulation parameters such that real-world conditions fall within the distribution of the training data. Domain randomization has proven effective in both simulated benchmarks and real-world robotic tasks when the data diversity is sufficiently large [4, 5, 6]. A seemingly unrelated but conceptually similar approach to visual generalization in the age of foundation models has been to train large
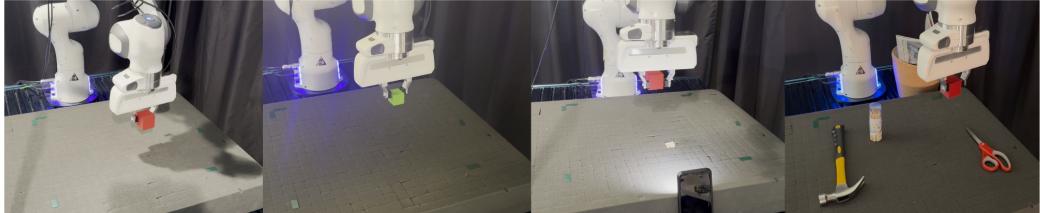
Figure 1: Behavior cloning with disentangled representations and associative latent dynamics achieves zero-shot generalization to various real world perturbations, such as changes in ambient lighting (**left**), object color (**middle-left**), directed lighting (**middle-right**), and the presence of distractor objects (**right**).

models, typically Vision-Language-Action (VLA) models, on very large real world robot datasets [7, 8, 9, 10]. By learning from varied contexts, these models generalize to novel environments, mirroring the principles of domain randomization at scale.

Despite recent advances, whether these models are truly capable of *extrapolative generalization* remains questionable. For example, [11] showed that vision-language models trained for pixel-wise future prediction perform poorly on out-of-distribution (OOD) data. Similarly, [12] found that state-of-the-art classifiers suffered up to a 30% accuracy drop when tested on objects in unusual poses. While one might mitigate this with orientation randomization, doing so would require exhaustively covering every possible orientation of every object we wish to classify – a clearly impractical solution. Compounding the issue, [13] shows that neural networks trained via supervised learning often fail to extract and generalize fundamental symmetries such as SO(2) rotations. This suggests that, despite all the efforts on exhaustive visual domain randomization in simulation or collecting massive and diverse real-world datasets, current models may not truly generalize and that data scaling, while important, is insufficient. Despite these shortcomings, it is well known that biological systems robustly extract structure from high-dimensional sensory inputs, even under severe visual perturbations [14, 15]. The neuroscience literature points to *factorized, modular* representations as a key enabler of this kind of structural generalization [15, 16, 17]. In machine learning, similar principles appear in disentangled and object-centric representation learning [18, 19, 20], which have been shown to facilitate visual generalization in continuous control tasks in simulation [21, 22, 23]. However, these approaches have yet to scale beyond toy problems or narrow benchmarks.

A recent approach that shows promise in scaling to harder tasks is Associative Latent DisentAnglement (ALDA) [23], a reinforcement learning (RL) algorithm that learns factorized representations via disentanglement and leverages principles of *Associative Memory* to achieve SOTA performance on a popular continuous control and visual generalization benchmark [24]. Given an OOD observation at test time, ALDA decomposes the observation into a disentangled latent representation and maps back *specific* dimensions that are OOD to in-distribution values, equivalent to recalling the most related observation it *has* seen and taking an action based on that instead. This work investigates whether ALDA's principles can extend to robotic manipulation and real-world deployment, where visual generalization remains challenging. While ALDA shows promise, RL alone struggles with complex manipulation tasks and cannot be directly trained from real-world interaction due to poor sample efficiency, making imitation learning a more practical alternative. We therefore extend ALDA to imitation learning, specifically diffusion-based behavior cloning methods [25, 26, 27], and find similar gains in zero-shot visual generalization. In addition to visual variations, real-world deployments can face camera perturbations, which further degrade performance. To address this, we draw on Equivariant Neural Networks, which encode symmetry structures (e.g., rotations) into their architecture and have shown strong generalization and sample efficiency in robot learning [28, 29] on robot learning tasks. However, these methods require training from scratch, significantly increasing training time and imposing constraints on the available model architectures. As an alternative, we introduce *learned canonicalization*, adapting recent work [30, 31] to the robot learning context. This family of methods uses a lightweight, surrogate equivariant neural network to *finetune* larger pre-

trained models to become equivariant to certain symmetries (e.g., SO(2) transformations) – without retraining from scratch. However, these methods have only been studied within simpler supervised learning contexts such as image classification and segmentation. We adapt this approach to robot learning and present an algorithm for turning any pre-trained robot policy into an equivariant one immune to discrete planar rotations in the SO(2) group.

To summarize, our contributions are as follows: **(1)** We evaluate ALDA on visually rich and dynamically complex manipulation tasks and demonstrate strong visual generalization, **(2)** we extend ALDA to imitation learning, specifically diffusion-based behavior cloning, and demonstrate similar gains in generalization performance over SOTA imitation learning baselines, **(3)** we propose a finetuning method to make any pretrained robot policy equivariant to discrete planar rotations, and **(4)** we validate our model on a real robot under realistic visual perturbations. By integrating these methods, we take a significant step towards generalist real-world agents capable of robust, zero-shot adaptation across lighting changes, background clutter, and camera perturbations.

2 Related Work

Reinforcement Learning for Robotics. RL algorithms learn a policy that maximizes the discounted sum of future rewards according to a given reward function. On-policy methods [32, 33, 34] learn from experience collected by the behavior policy, while off-policy methods [35, 36, 37, 38] can learn from prior experience. Model-based RL algorithms construct explicit predictive models of the environment, which are used for planning, and can learn from imagined experience [39, 40, 41, 42]. These methods have been shown to be effective at solving tabletop manipulation tasks [41], contact-rich assembly tasks [43, 44], and dexterous manipulation [4, 45] on modern simulators [1, 2, 46] and on real robots.

Learning from Demonstrations. Due to the sample inefficiency of RL and the complexity of certain tasks, imitation learning has become an increasingly popular paradigm that allows robots to leverage sparser real-world, expert demonstration datasets [47, 48]. Generative methods such as diffusion models [49, 50, 51, 52] have shown success at solving manipulation tasks where the data distribution exhibits multiple modalities [25, 26, 27, 53, 54]. Recently, Transformer [55]-based methods have shown dramatic improvements in generalization across tasks [56, 57, 58]. In a similar vein, there has been a recent trend to train "Robot Foundation Models", using either from-scratch Transformer models or pre-trained Vision-Language Models (VLMs) converted into Vision-Language-Action (VLA) models by training on large demonstration datasets [8, 9, 7, 10]. However, there is a growing body of evidence [13, 12, 11] to suggest that LLMs, VLMs, and more broadly current Deep Learning architectures, are not achieving systematic extrapolative generalization, implying that the downstream models finetuned for robotics tasks likely have gaps in performance and many edge cases despite claims of generalization.

Representation Learning for Control. We concur with and adopt the position taken by several recent works [11, 59, 60] which argue that structured representations are crucial to extrapolative generalization in agentic tasks like future prediction and planning. From the cognitive and neuroscience literature, there is mounting evidence that biological agents are capable of rapid adaptation and generalization in part thanks to modular, structured representations [15, 14, 17, 61, 62], which we posit will greatly aid artificial agents in doing the same. One such approach to learning structured representations in machine learning is through disentangled representation learning (DRL) [18, 63, 19, 64], in which a model learns a latent representation of high-dimensional data (e.g., images), where each dimension represents one factor of variation (e.g., object color, size, shape, etc.). DRL is thought to be a key component of compositional generalization [65, 15], and has shown promising results in visual generalization in continuous control tasks from high-dimensional image observations [23, 22, 21]. ALDA [23], which we use as the foundation for this work, combines learning factorized representations with principles of associative memory from modern, continuous Hopfield Networks [66, 67, 68]. Hopfield networks store memories as fixed points and use attractor dynamics to recover the most similar memory given an input query. Recent
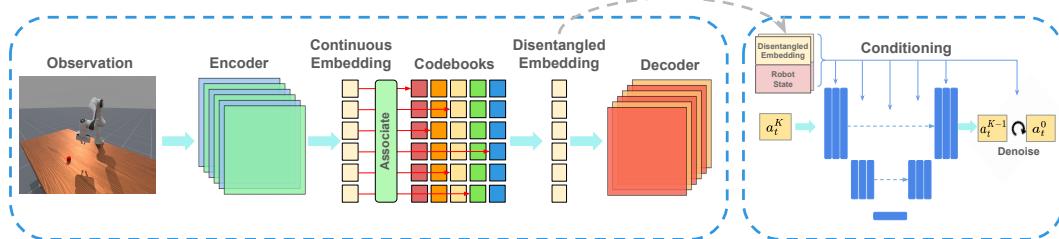
Figure 2: Overview of ALDA + Diffusion Policy (**ALDA-DP**). ALDA-DP jointly learns a factorized representation of the image observation while training the policy. The diffusion model denoises actions conditioned on this representation.

work in neuroscience finds evidence that the hippocampus in mice achieves rapid generalization through disentangled memory representations [17], providing a biologically plausible motivation for this line of work.

Equivariance via Learned Canonicalization. Learned canonicalization is a method by which a larger pretrained network, called the predictor network p, can be made equivariant using a smaller, surrogate equivariant neural network. An equivariant function h is one that commutes with a *group action* $\rho(g)$ on a symmetry group $\mathcal{G}$ such that $h(\rho(g) \cdot \mathbf{o}) = \rho'(g) \cdot h(\mathbf{o})$ i.e. an equivariant transformation on the input $\mathbf{o}$ induced by $\mathcal{G}$ results in a predictable transformation of the output $h(\mathbf{o})$. Normally, the predictor network would need to learn the inverse mapping from h. However, one can instead learn a *canonicalization network* $C : \mathcal{O} \to G$ that can transform elements of the input from their orbit into a *canonical* sample where h can be applied, and then transforming the sample back to its original position in the orbit:

$$h(\mathbf{o}) = \rho'(C(\mathbf{o})) \cdot p \left(\rho(C(\mathbf{o})^{-1}) \cdot \mathbf{o} \right) \tag{1}$$

This formulation alleviates the burden of modifying the pretrained network p, instead putting it on the canonicalization network, which can be a lightweight equivariant neural network. This can be especially useful when training robot policies where issues such as stabilizing learning and sample/time complexity are more relevant. Indeed, [28] requires using equivariant neural networks for the RL agent itself, along with additional assumptions on the MDP that ensure SO(2) equivariance is not violated during training. Learned canonicalization has shown success on image classification and segmentation tasks [30, 31], but has not been studied in the robot learning context. To the best of our knowledge, we are the first to demonstrate how this method can be used to make vision-based robot policies robust to camera perturbations.

3   Method

Diffusion models have shown impressive results on current manipulation benchmarks [26, 27, 25, 54, 53]; thus, we choose a diffusion-based actor as our driver. We describe our approach to learning disentangled representations and leveraging principles of associative memory for Diffusion Policy [25]. An overview of the method is presented in Figure 2. From now on, we will refer to the Diffusion Policy variant as ALDA-DP, and the RL variant as ALDA-SAC. ALDA-SAC learns a factorized representation while jointly training an RL agent using Soft Actor-Critic [35], and assumes the same observation and action representations. We refer the reader to [23] and the Appendix for implementation details on ALDA-SAC.

3.1   Disentangling Representations for Behavior Cloning

ALDA receives an image observation and encodes it into a continuous latent representation $z_{cont.} \in \mathbb{R}^{n_z}$ with the encoder f_θ. $z_{cont.}$ is mapped to a discrete representation $z_d \in \mathbb{R}^{n_z}$ via a *collection* of discrete scalar codebooks $Z = V_1 \times V_2 \times ... \times V_{n_z}$, one per latent dimension. The mapping is given

by an associative latent dynamics model, which leverages the attention mechanism used in modern Hopfield networks to perform *pointwise* attention between scalar values in $z_{cont.}$ and the codebooks to produce z_d:

$$z_{d_j} = \text{Softmax}\left(\beta\text{Sim}(z_j, V_j)\right) \odot V_j \tag{2}$$

A decoder network g_ϕ uses z_d to reconstruct the observation, thus propagating visual information back to z_d. The discrete nature of z_d enforces separation and, combined with large network activation penalties, facilitates disentanglement within z_d. More details about the associative latent mechanism and disentanglement can be found in the Appendix. Of particular importance is that when presented with in-distribution observations at test time, the association step in (2) is approximately a no-op, because the encoder has been optimized to be close to the distribution $\mathbf{z}_d$ and thus $\mathbf{z}_{cont.}$ will likely already be $\mathbf{z}_d$ or at least very similar. However, if the observations become out of distribution due to visual distractions, then Equation (2) *forces* the representation to be in-distribution before the actor model sees it. Unlike tasks with learnable modern Hopfield networks where the associations between two sets need to be learned explicitly, here, association is made possible without learning *because* the representations are disentangled.

3.2 Training

Diffusion-based policies learn to iteratively denoise actions conditioned on latent representations z_t of image observations $\mathbf{o} \in \mathcal{O}$ and robot proprioceptive state $\mathbf{s} \in \mathcal{S}$. Following prior works [25, 26], we assume access to and learn from a dataset of expert action trajectories $\{(\mathbf{o}_1, \mathbf{s}_1, \mathbf{a}_1), (\mathbf{o}_2, \mathbf{s}_2, \mathbf{a}_2), ..., (\mathbf{o}_T, \mathbf{s}_t, \mathbf{a}_T)\}$, where $\mathbf{o}_t$ is an image observation, $\mathbf{s}_t$ robot proprioceptive state, and $\mathbf{a}_t$ is a robot action. Actions $a_t \in \mathbb{R}^4$ are delta *xyz* position commands for the end-effector and a gripper open/close command $a_t = \{a_t^{\Delta_{\text{loc}}}, a_t^{\text{open}} \in \{0, 1\}\}$. The actor predicts subsequences of future actions $\tau = a_{t:t+k}$ to a horizon of length k. We use a denoising probabilistic diffusion model [49], which is trained by iteratively adding noise according to a variance schedule β_i to action subsequences and learning the inverse denoising procedure $p_\theta(\tau^{i-1}|\tau^i) = \mathcal{N}(\tau^{i-1}; \mu_\theta(\tau^i, i), \Sigma_\theta(\tau^i, i))$. During inference, the diffusion model, conditioned on the disentangled latent $\mathbf{z}_d$ and robot proprioceptive state $\mathbf{s}$, denoises a noisy action chunk τ^i into a noise-free action sequence τ^0 that is executed by the robot. During training, we randomly sample a trajectory timestep t and diffusion timestep i and add noise ϵ to the ground truth action sequence τ^0. We use mean-squared error (MSE) to predict the noise at i:

$$J(DP) = ||\epsilon_\theta(\mathbf{z}_d, \mathbf{s}, \tau^i, i) - \epsilon||_2^2. \tag{3}$$

The final training objective is therefore $J(ALDA) + J(DP)$ (see the Appendix for details on $J(ALDA)$).

3.3 Equivariant Adaptation

We now describe our technique to make pre-trained policies invariant to discrete image rotations in SO(2), based on the method presented in [30]. Assume we are given a trained policy $\pi(a|z)$, encoder f, and latent model l. An lightweight equivariant neural network is initialized as the canonicalization function C. To enable our policy to take optimal actions under camera rotations, we must make f and l equivariant to group actions on $\mathbf{o}$ *and* make π equivariant to the resulting group actions on z. Therefore, the prediction function p in equation (1) we wish to optimize is $\pi(\cdot|l(f(\mathbf{o})))$. The canonicalization method presented in [30] was studied under the context of supervised learning and assumes the existence of a ground truth dataset. In our case, the trained actor and encoder networks and latent model are our oracles, so we duplicate and freeze their weights, and refer to them as π^*, f^*, and l^*. Rather than a reconstruction loss over images as in prior works, we wish to "reconstruct" the optimal actions of $\pi^*(\cdot|f^*(\cdot))$ given the canonicalized sample $C(\mathbf{o})$ of the original observation $\mathbf{o}$. Since our RL variant of ALDA under the hood is Soft Actor-Critic, which maintains a replay buffer, we can save and reuse the buffer as a dataset to sample i.i.d. transitions. For the ALDA-DP variant, we sample transitions from the expert demonstration dataset.

Since we disable gradient flow through the latent model, we add the commitment loss term from ALDA's training objective here to keep the continuous outputs of f and discrete embeddings of l close to each other during the optimization procedure. Finally, following in step with prior work, we also utilize a canonicalization prior (CP) regularizer to ensure the canonicalized inputs match the original observation inputs as closely as possible. This is done by minimizing the KL-Divergence between the transformed distribution induced by C and the original data distribution i.e. $\mathcal{L}_{prior} = -\mathbb{E}_{\mathbf{o} \sim D}\left[D_{KL}(\mathbb{P}_D || P_{C(\mathbf{o})})\right]$. The final objective can then be written as

$$||\pi(a|l(f(\mathbf{o})) - \pi^*(a|l^*(f^*(\mathbf{o}))||_2^2 + \beta \cdot \mathcal{L}_{prior} + \mathcal{L}_{commit} \quad (4)$$

where β is a hyperparameter controlling the regularization strength. Since prior learned canonicalization methods have struggled with continuous image rotations, we also restrict our approach to the space of discrete rotations C_n. However, the number of discrete bins n can be increased for robustness to more degrees of rotational camera perturbations, albeit at the expense of computational complexity. Pseudocode for the finetuning procedure is in the Appendix.

4 Experiments

Figure 3: ManiSkill3 visual generalization tasks. **Left to right**: random lighting, random cube color, distracting background (DBG), DBG + random cube color, DBG + random lighting + random cube color.

To investigate how our method scales to manipulation tasks, we choose ManiSkill3's [1] tabletop manipulation suite. ManiSkill3 is a high-throughput simulator with support for GPU parallelization. It comes with demonstrations for imitation learning on many tabletop tasks out of the box, making it an obvious choice for benchmarking. Since we wish to test the robustness of our method and baselines to visual distribution shifts, we design a suite of visual scene randomizations on top of existing tasks, which we call the **ManiSkill Visual Generalization Benchmark** (MVGB). MVGB supports background scene randomizations, lighting intensity and lighting direction randomization, table color randomization, and object color/size randomization. For this work, we benchmark on six sources of visual randomization, which are combinations of the following three principle randomizations: **(i) Distracting Backgrounds (DBG)** where we overlay an image randomly sampled from the Places365 dataset [69] consisting of 1.8 million images from 365 different real-world environments, **(ii) Random Colors**, where the color of the object being manipulated is randomized, and **(iii) Random Lighting**, where the scene's ambient lighting color and intensity is randomized.

We compare ALDA-SAC and ALDA-DP against various SOTA RL and BC baselines, respectively. For ALDA-SAC, we compare to **SAC**, **SAC-AE** [21], which demonstrated promising generalization performance by training an autoencoder as an auxiliary objective, and **TD-MPC2** [41], a SOTA model-based RL algorithm that uses Model Predictive Control (MPC) to plan in the learned latent space of the model. Results are presented on three tasks – PickCube, PushCube, and PullCube. For ALDA-DP, we compare against **Diffusion Policy** [25] and **Action Chunking with Transformers** [58] (ACT) on PickCube, PushCube, and the more difficult, long-horizon task PushT. All methods are evaluated on the six visual variations shown in Figure 3 for their respective tasks. For each variation, we compute the average success rate over 1000 rollouts for the RL methods. Due to computational constraints, we computed success rates over 500 rollouts for the BC methods. The results are presented in Figure 4.
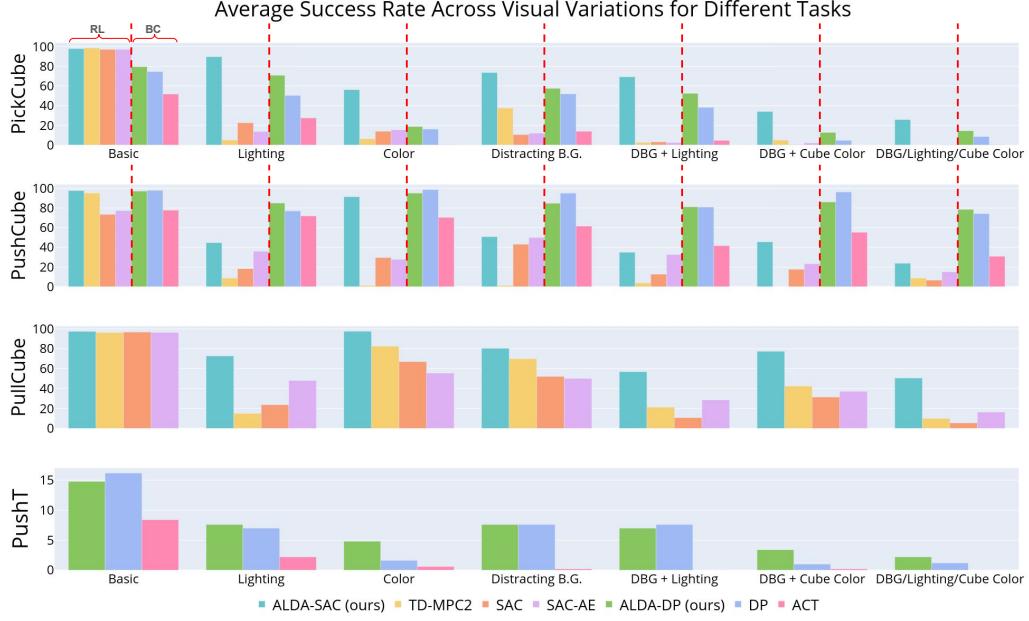
Figure 4: Average success rate of ALDA-SAC and ALDA-DP compared to various RL and BC baselines. The first two rows have both RL and BC results, while the 3rd row has only RL results and the fourth only BC results. ALDA-* methods overall perform the best, and with a large margin, especially on PickCube.

On PickCube, ALDA-SAC performs the best by significant margins on all visual variations, followed by ALDA-DP. On PushCube and PullCube, ALDA-SAC once again outperforms RL baselines by large margins. BC methods tend to outperform their RL counterparts on PushCube, where ALDA-DP and DP perform roughly the same, followed by ACT. We suspect that a combination of action-chunking, i.e., executing a sequence of actions before observing the next state, combined with the simplicity of the non-prehensile pushing task, makes unaltered Behavior Cloning baselines fairly robust to visual perturbations without requiring structured representations or associative memory. However, when fine-grained and precise manipulation is required, as with PickCube, visual randomizations present more of a challenge and are where ALDA-DP demonstrates superior performance. The performance across ALDA- methods and baselines on PushT is low even without visual perturbations. Nonetheless, we find that ALDA-DP outperforms baselines on most visual variations, suggesting that a future, more powerful underlying BC algorithm that performs better on the base task will benefit from the structured representations that ALDA provides.

4.1 Equivariant Adaptation Results

We finetune ALDA-DP and ALDA-SAC using the equivariant adaptation technique (Section 3.3) on three discrete cyclic groups, C_8, C_{12}, and C_{24}, i.e., three separate finetuned models. These correspond to image rotations in 45, 30, and 15 degree increments, respectively. We train both models for 500 iterations, which takes at most 7 minutes for ALDA-DP on C_{24}, and 15 minutes for ALDA-SAC on C_{24} on an RTX 3090. Evaluation on C_n implies n different image rotations, and for each rotation, we evaluate the finetuned model over 100 parallel environments and compute the average success rate. The average over all rotations' success rates on the PickCube task is in Table 1.

ALDA-SAC performs the best and maintains high success rates, even with larger values of N. While ALDA-DP benefits from the finetuning procedure, we notice a drop-off in performance as the rotational discretization becomes more fine-grained. We hypothesize ALDA-DP is more challenging to finetune due to the larger model size, and will likely benefit from more iterations and/or deeper canonicalization networks. Nonetheless, our finetuning procedure results in strong robustness to

Image Rotations	ALDA-SAC	ALDA-SAC (no finetune)	ALDA-DP	ALDA-DP (no finetune)
None	98.00	98.00	79.69	79.69
C_8 (Δ45 degrees)	**97.44**	1.78	**76.86**	3.79
C_{12} (Δ30 degrees)	**96.12**	1.71	**60.94**	4.19
C_{24} (Δ15 degrees)	**96.05**	1.60	**45.57**	3.16

Table 1: Results of ALDA models finetuned using the equivariant adaptation technique on the Pick-Cube task. For Group n, image observations are rotated every $\frac{360}{n}$ degrees, and success rates are calculated across 100 parallel environments. The average success rate over all rotations in the cyclic group is presented here.

image rotations, which will significantly aid in mitigating the effects of camera perturbations during real-world deployment.

4.2   Real-World Experiments

Algorithm	Basic	Directed Light (Left/Middle/Right)	Ambient Light	Gray Cube	Distracting Objects
ALDA Diffusion Policy	80.0	(**70.0** / 0.0 / 0.0)	**85.0**	**30.0**	**60.0**
Diffusion Policy	80.0	(0.0 / 0.0 / 0.0)	0.0	0.0	0.0
ACT	80.0	(55.0 / 0.0 / 0.0)	15.0	0.0	0.0

Table 2: Average success rate of ALDA-DP, DP, and ACT on the PickCube task with the Franka Emika Panda arm under various visual perturbations. Results are averaged over 20 trials.

We collected 200 demonstrations via teleoperation of the Franka Emika Panda arm picking up a red cube, and used this dataset to train ALDA-DP, DP, and ACT. For evaluating the models, we designed 6 visual perturbation experiments and recorded the average success rate of each method over 20 trials. The "directed light" perturbations involve shining a flashlight on the workspace at three different angles – from the left, middle, and right of the workspace. For the "ambient light" randomization, we turn on the overhead lights in the workspace, resulting in shadows of the arm and tabletop objects being cast onto the table. To test robustness to color randomizations, we tried picking up a gray cube instead of a red one. Finally, for the "distracting objects" randomization, we place various objects in the workspace and the background. Visualizations of these perturbations are presented in Figure 1, and the results are presented in Table 2. With the exception of ACT achieving decent performance on Directed Light (Left), ALDA-DP outperforms all baselines by significant margins. Directed Light Middle and Right proved to be too difficult for ALDA-DP, since the flashlight beam can cause visual occlusions of the cube and change its perceived color.

5   Conclusion

We present strong empirical evidence that disentangled representations, when paired with associative latent dynamics, enable robust zero-shot visual generalization for complex manipulation tasks in both simulation and real-world settings. Our model achieves this without data augmentation, domain randomization, or camera calibration, and is robust to certain camera perturbations. Outperforming SOTA RL and BC baselines, this approach marks meaningful progress toward generalist agents capable of solving tasks in the wild and provides an alternative to typical generalization paradigms to the robot learning community. While the work here makes significant strides, much remains to be done in improving generalization performance. For one, data diversity and data scaling performance are without question important pillars of developing generalist agents, and it remains to be seen if and how these techniques can be scaled to large models on massive datasets. Nonetheless, we believe the work here is a step in the right direction, and the remaining challenges present exciting opportunities for further research.

6   Limitations

We attempted several visual perturbations for which our approach was unsuccessful. In simulation, changing the table color at evaluation time led to a near-total performance collapse of the model. We suspect that without training data where the table and object colors are randomized independently,

it is possible that the latent disentanglement procedure fails to separate object color and table color as independent sources of variation. Thus, changing one during evaluation can change the other and affect the downstream performance of the policy. Indeed, randomizing the table color during training mitigates the performance collapse, validating that data diversity is still an important component even when learning structured representations. During real-world evaluations, lifting the black curtains in the background (see Figure 1) causes sunlight to bleed into the camera frame, resulting in an over-saturated white background that also causes performance collapse. While ALDA agents demonstrate robustness to distracting backgrounds, it seems that certain lighting conditions, as we also noticed with directed middle and right lighting (Table 2), remain challenging for our method. This could perhaps be mitigated with camera exposure balancing, and we remain optimistic that, with the right amount of exposure tuning, our method will zero-shot adapt to other scenes, such as other labs and rooms. Nonetheless, these negative results show that there is still much room for improvement.

Acknowledgments

References

[1] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/eda80a3d5b344bc40f3bc04f65b7a357-Abstract-round2.html.

[2] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. Isaac gym: High performance GPU based physics simulation for robot learning. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/28dd2c7955ce926456240b2ff0100bde-Abstract-round2.html.

[3] P. Katara, Z. Xian, and K. Fragkiadaki. Gen2sim: Scaling up robot learning in simulation with generative models. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pages 6672–6679. IEEE, 2024. doi:10.1109/ICRA57147.2024.10610566. URL https://doi.org/10.1109/ICRA57147.2024.10610566.

[4] M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba. Learning dexterous in-hand manipulation. *Int. J. Robotics Res.*, 39(1), 2020. doi:10.1177/0278364919887447. URL https://doi.org/10.1177/0278364919887447.

[5] A. Almuzairee, N. Hansen, and H. I. Christensen. A recipe for unbounded data augmentation in visual reinforcement learning. *RLJ*, 1:130–157, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper26.html.

[6] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand. *CoRR*, abs/1910.07113, 2019. URL http://arxiv.org/abs/1910.07113.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model. In P. Agrawal, O. Kroemer, and W. Burgard, editors, *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR, 2024. URL https://proceedings.mlr.press/v270/kim25c.html.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. T. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. RT-1: robotics transformer for real-world control at scale. In K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi:10.15607/RSS.2023.XIX.025. URL https://doi.org/10.15607/RSS.2023.XIX.025.

[9] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. T. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski,

Y. Lu, S. Levine, L. Lee, T. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In J. Tan, M. Toussaint, and K. Darvish, editors, *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

[10] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine. FAST: efficient action tokenization for vision-language-action models. *CoRR*, abs/2501.09747, 2025. doi:10.48550/ARXIV.2501.09747. URL https://doi.org/10.48550/arXiv.2501.09747.

[11] A. Nayebi, R. Rajalingham, M. Jazayeri, and G. R. Yang. Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html.

[12] A. Abbas and S. Deny. Progress and limitations of deep networks to recognize objects in unusual poses. In B. Williams, Y. Chen, and J. Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 160–168. AAAI Press, 2023. doi:10.1609/AAAI.V37I1.25087. URL https://doi.org/10.1609/aaai.v37i1.25087.

[13] A. Perin and S. Deny. On the ability of deep networks to learn symmetries from data: A neural kernel theory. *CoRR*, abs/2412.11521, 2024. doi:10.48550/ARXIV.2412.11521. URL https://doi.org/10.48550/arXiv.2412.11521.

[14] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[15] T. E. Behrens, T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

[16] J. J. Bakermans, J. Warren, J. C. Whittington, and T. E. Behrens. Constructing future behavior in the hippocampal formation through composition and replay. *Nature Neuroscience*, pages 1–12, 2025.

[17] W. Tang, H. Chang, C. Liu, S. Perez-Hernandez, W. Y. Zheng, J. Park, A. Oliva, and A. Fernandez-Ruiz. A hippocampal population code for rapid generalization. *bioRxiv*, pages 2025–03, 2025.

[18] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[19] K. Hsu, W. Dorrell, J. C. R. Whittington, J. Wu, and C. Finn. Disentanglement via latent quantization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine,

editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8e63972d4d9d81b31459d787466ce271-Abstract-Conference.html.

[20] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf. Object-centric learning with slot attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

[21] D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau, and R. Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10674–10681. AAAI Press, 2021. doi:10.1609/AAAI.V35I12.17276. URL https://doi.org/10.1609/aaai.v35i12.17276.

[22] I. Higgins, A. Pal, A. A. Rusu, L. Matthey, C. P. Burgess, A. Pritzel, M. M. Botvinick, C. Blundell, and A. Lerchner. DARLA: improving zero-shot transfer in reinforcement learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1480–1490. PMLR, 2017. URL http://proceedings.mlr.press/v70/higgins17a.html.

[23] S. Batra and G. S. Sukhatme. Zero-shot generalization of vision-based RL without data augmentation. *CoRR*, abs/2410.07441, 2024. doi:10.48550/ARXIV.2410.07441. URL https://doi.org/10.48550/arXiv.2410.07441.

[24] N. Hansen, H. Su, and X. Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. 2021.

[25] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi:10.15607/RSS.2023.XIX.026. URL https://doi.org/10.15607/RSS.2023.XIX.026.

[26] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In D. Kulic, G. Venture, K. E. Bekris, and E. Coronado, editors, *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi:10.15607/RSS.2024.XX.067. URL https://doi.org/10.15607/RSS.2024.XX.067.

[27] T. Ke, N. Gkanatsios, and K. Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In P. Agrawal, O. Kroemer, and W. Burgard, editors, *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pages 1949–1974. PMLR, 2024. URL https://proceedings.mlr.press/v270/ke25a.html.

[28] D. Wang, R. Walters, and R. Platt. $\mathrm{SO}(2)$-equivariant reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=7F9cOhdvfk_.

[29] K. Chen, X. Chen, Z. Yu, M. Zhu, and H. Yang. Equidiff: A conditional equivariant diffusion model for trajectory prediction. In *25th IEEE International Conference on Intelligent Transportation Systems, ITSC 2022, Macau, China, October 8-12, 2022*, pages 746–751. IEEE, 2023. doi:10.1109/ITSC57777.2023.10421892. URL https://doi.org/10.1109/ITSC57777.2023.10421892.

[30] A. K. Mondal, S. S. Panigrahi, O. Kaba, S. Mudumba, and S. Ravanbakhsh. Equivariant adaptation of large pretrained models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9d5856318032ef3630cb580f4e24f823-Abstract-Conference.html.

[31] S. Kaba, A. K. Mondal, Y. Zhang, Y. Bengio, and S. Ravanbakhsh. Equivariance with learned canonicalization functions. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15546–15566. PMLR, 2023. URL https://proceedings.mlr.press/v202/kaba23a.html.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

[33] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz. Trust region policy optimization. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/schulman15.html.

[34] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In M. Balcan and K. Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1928–1937. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/mniha16.html.

[35] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1856–1865. PMLR, 2018. URL http://proceedings.mlr.press/v80/haarnoja18b.html.

[36] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.02971.

[37] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1582–1591. PMLR, 2018. URL http://proceedings.mlr.press/v80/fujimoto18a.html.

[38] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Ried-miller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

[39] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

[40] D. Hafner, T. P. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR, 2019. URL http://proceedings.mlr.press/v97/hafner19a.html.

[41] N. Hansen, H. Su, and X. Wang. TD-MPC2: scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

[42] D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[43] B. Tang, M. A. Lin, I. Akinola, A. Handa, G. S. Sukhatme, F. Ramos, D. Fox, and Y. S. Narang. Industreal: Transferring contact-rich assembly tasks from simulation to reality. In K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi:10.15607/RSS.2023.XIX.039. URL https://doi.org/10.15607/RSS.2023.XIX.039.

[44] B. Tang, I. Akinola, J. Xu, B. Wen, A. Handa, K. V. Wyk, D. Fox, G. S. Sukhatme, F. Ramos, and Y. S. Narang. Automate: Specialist and generalist assembly policies over diverse geometries. In D. Kulic, G. Venture, K. E. Bekris, and E. Coronado, editors, *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi:10.15607/RSS.2024.XX.064. URL https://doi.org/10.15607/RSS.2024.XX.064.

[45] A. Handa, A. Allshire, V. Makoviychuk, A. Petrenko, R. Singh, J. Liu, D. Makoviichuk, K. V. Wyk, A. Zhurkevich, B. Sundaralingam, and Y. S. Narang. Dextreme: Transfer of agile in-hand manipulation from simulation to reality. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 5977–5984. IEEE, 2023. doi:10.1109/ICRA48891.2023.10160216. URL https://doi.org/10.1109/ICRA48891.2023.10160216.

[46] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. P. Lillicrap, and M. A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018. URL http://arxiv.org/abs/1801.00690.

[47] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In A. Faust, D. Hsu, and G. Neumann, editors, *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 1678–1690. PMLR, 2021. URL https://proceedings.mlr.press/v164/mandlekar22a.html.

[48] S. Haldar, J. Pari, A. Rai, and L. Pinto. Teach a robot to FISH: versatile imitation from one minute of demonstrations. In K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi:10.15607/RSS.2023.XIX.009. URL https://doi.org/10.15607/RSS.2023.XIX.009.

[49] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[50] Y. Zeng, M. Suganuma, and T. Okatani. Inverting the generation process of denoising diffusion implicit models: Empirical evaluation and a novel method. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pages 4516–4524. IEEE, 2025. doi:10.1109/WACV61041.2025.00453. URL https://doi.org/10.1109/WACV61041.2025.00453.

[51] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi:10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

[53] X. Zhang, M. Chang, P. Kumar, and S. Gupta. Diffusion meets dagger: Supercharging eye-in-hand imitation learning. In D. Kulic, G. Venture, K. E. Bekris, and E. Coronado, editors, *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi:10.15607/RSS.2024.XX.048. URL https://doi.org/10.15607/RSS.2024.XX.048.

[54] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann, and S. Devlin. Imitating human behaviour with diffusion models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Pv1GPQzRrC8.

[55] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[56] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In K. Liu, D. Kulic, and J. Ichnowski, editors, *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 785–799. PMLR, 2022. URL https://proceedings.mlr.press/v205/shridhar23a.html.

[57] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y. Chao, and D. Fox. RVT: robotic view transformer for 3d object manipulation. In J. Tan, M. Toussaint, and K. Darvish, editors, *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 694–710. PMLR, 2023. URL https://proceedings.mlr.press/v229/goyal23a.html.

[58] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi:10.15607/RSS.2023.XIX.016. URL https://doi.org/10.15607/RSS.2023.XIX.016.

[59] Q. Garrido, N. Ballas, M. Assran, A. Bardes, L. Najman, M. Rabbat, E. Dupoux, and Y. LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *CoRR*, abs/2502.11831, 2025. doi:10.48550/ARXIV.2502.11831. URL https://doi.org/10.48550/arXiv.2502.11831.

[60] J. Pari, N. M. M. Shafiullah, S. P. Arunachalam, and L. Pinto. The surprising effectiveness of representation learning for visual imitation. In K. Hauser, D. A. Shell, and S. Huang, editors, *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, 2022. doi:10.15607/RSS.2022.XVIII.010. URL https://doi.org/10.15607/RSS.2022.XVIII.010.

[61] V. Samborska, J. L. Butler, M. E. Walton, T. E. Behrens, and T. Akam. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience*, 25(10):1314–1326, 2022.

[62] W. Sun, M. Advani, N. Spruston, A. Saxe, and J. E. Fitzgerald. Organizing memories for generalization in complementary learning systems. *Nature neuroscience*, 26(8):1438–1448, 2023.

[63] J. C. R. Whittington, W. Dorrell, S. Ganguli, and T. Behrens. Disentanglement with biological constraints: A theory of functional cell types. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=9Z_GfhZnGH.

[64] K. Hsu, J. I. Hamid, K. Burns, C. Finn, and J. Wu. Tripod: Three complementary inductive biases for disentangled representation learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=0iXp5P77ho.

[65] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P. Burgess, M. Bosnjak, M. Shanahan, M. M. Botvinick, D. Hassabis, and A. Lerchner. SCAN: learning hierarchical compositional visual concepts. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rkN2Il-RZ.

[66] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[67] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, T. Adler, D. P. Kreil, M. K. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

[68] B. Hoover, Y. Liang, B. Pham, R. Panda, H. Strobelt, D. H. Chau, M. Zaki, and D. Krotov. Energy transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

[69] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
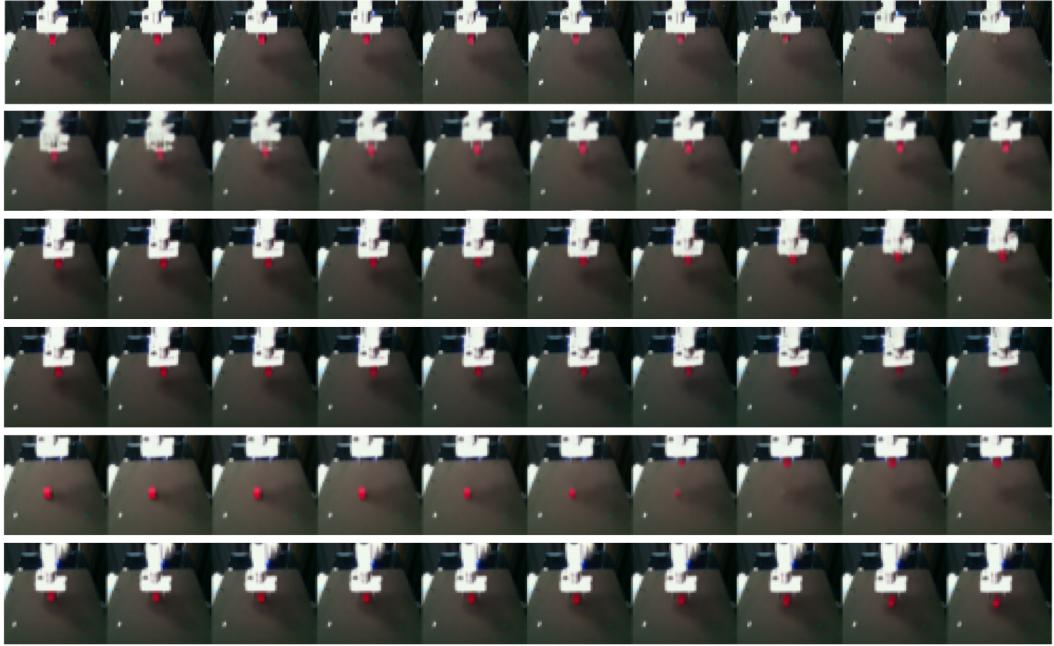
A   Latent Traversals

Figure 5: Traversing the disentangled latent space of ALDA-DP trained on real demonstrations and visualizing their corresponding reconstructions. Rows correspond to latent traversals of a single latent dimension given a reference image. By editing specific latent dimensions, we can visualize what factor of variation they correspond to.

Since there are no quantitative metrics that tell us how well the representation disentangles without knowing the ground truth sources of variation a priori, following [23], we present qualitative results instead. In this "latent traversal" experiment, we sample a batch of image observations from the expert-demonstrations dataset collected on the real Franka arm and encode them into disentangled representations using a trained ALDA-DP model. From here, we randomly sample an image from the batch and edit a randomly chosen dimension of the disentangled representation, interpolating it from [-1, 1]. The modified latent code is then passed to the decoder for visualizing the reconstructed image. Each row corresponds to the resulting reconstructed images from editing and interpolating a single latent dimension given a reference observation. The different rows correspond to latent traversals of different reference images. This experiment allows us to qualitatively see if editing a single latent dimension corresponds to a singular change in the resulting image, while also determining semantically what factor of variation it learns to represent.

From Figure 5, we find that indeed one latent dimension seems to correspond to a single factor of variation. For example, in the last row, the traversal of this latent dimension corresponds to different cube positions along the x-axis. Interestingly, several of the latent traversals show discontinuities, such as the second to last row where the cube is either on the table or picked up by the gripper. We hypothesize two reasons for this: (1) the discrete nature of the disentangled latent representation results in discontinuities, and (2) the task gradients from the BC objective encourage the agent to only pay attention to the cube either when it has yet to be picked up, or when it is already picked up and moved to the goal position, since these "critical" states potentially matter more than intermediate states. Future research into associative latent disentanglement models with continuous representations and/or representations that are more temporally consistent will greatly aid in certain tasks such as those with contact-rich dynamics or that require precise control.

B Details on Associative Latent DisentAnglement

We first provide a formal definition of disentanglement. Since we wish to learn a representation such that each dimension of the latent embedding corresponds to a single factor of variation using a nonlinear model (e.g. neural network), the disentangled representation learning problem is often formulated as one of Nonlinear Independent Component Analysis, or **Nonlinear ICA**. Suppose we are given a dataset of images $\mathcal{D} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_N\}$ with n_s nonlinear independent source variables $s_1, ..., s_{n_s}$ that account for all variations within the data distribution. A hidden nonlinear generator function $g : \mathcal{S} \rightarrow \mathcal{O}$ maps the sources to the observations:

$$p(\mathbf{s}) = \prod_{i=1}^{n_s} p(s_i), \mathbf{o} = g(\mathbf{s}).$$

The goal of Nonlinear ICA and thus disentangled representation learning is to uncover the hidden sources $s_1, ..., s_{n_s}$ factorized from each other.

Under the hood, ALDA employs an adaptation of QLAE [19] to disentangle the latent representation. An image observation $\mathbf{o}_t$ is first encoded into a continuous representation $\mathbf{z}_{cont.} \in \mathbb{R}^{n_z}$, where n_z is the number of independent sources of variation that form the basis of the observation distribution. Each dimension $z_j, j = 1, ..., n_z$ of $\mathbf{z}_{cont.}$ is mapped to a discrete value by a collection of scalar codebooks $Z = V_1 \times V_2 \times ... \times V_{n_z}$, one per latent dimension via attention-based association:

$$z_{d_j} = \text{Softmax}\left(\beta \text{Sim}(z_j, V_j)\right) \odot V_j$$

where $\text{Sim}(\cdot, \cdot)$ is any similarity function and β is a hyperparameter that controls the degree of separation between latent values. We use the negative L_1 distance as our similarity function: The disentangled latent representation is used to reconstruct the original observation as an auxiliary objective using a reconstruction loss $\mathcal{L}_{recon.}$ Although gradients can flow through the Softmax operator, in practice and with large values of β, the gradients can become extremely large and destabilize training, so we instead use a StopGrad operator and optimize the encoder to be close to the discretized latent distribution using a commitment loss term:

$$\mathcal{L}_{commit.} = ||\text{StopGrad}(\mathbf{z}_{cont.}) - \mathbf{z}_d||_2^2.$$

Large activation penalties $\lambda_\theta ||\theta||_2^2, \lambda_\phi ||\phi||_2^2$ are applied to the encoder and decoder weights θ and ϕ serving as an information bottleneck which, inline with prior disentanglement methods [64, 18, 63], facilitates the disentanglement process. Put together, the ALDA objective is

$$J(ALDA) = \mathbb{E}_{\mathbf{o}_t \sim \mathcal{D}}\left[w_1 \mathcal{L}_{recon.} + w_2 \mathcal{L}_{commit} + \lambda_\theta ||\theta||_2^2 + \lambda_\phi ||\phi||_2^2\right].$$

For all experiments, we set λ_ϕ and λ_θ to 0.1 and w_1, w_2 to 1.0 and 0.1, respectively.
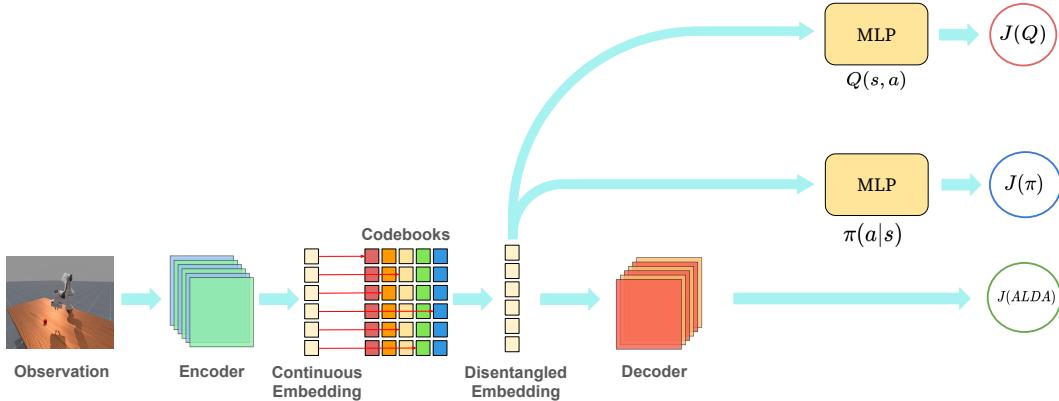
C ALDA-SAC

Figure 6: ALDA-SAC method diagram.

The original ALDA-SAC implementation in [23] assumes as input a stack of observations $\in \mathbb{R}^{k \times C \times H \times W}$. The ALDA model is trained to independently encode and reconstruct each individual frame and disentangle the latent representation according to the objective $J(ALDA)$ (Section B). The stack of latent states $\in \mathbb{R}^{k \times |z_d|}$ is fed into a 1D-CNN to extract temporal information before being input into a linear projection layer and then finally the actor/critic networks. However, we did not find frame stacking or projection to be beneficial on the ManiSkill3 manipulation tasks, and instead opt for a simplified implementation, illustrated in Figure 6. Here, a single observation is disentangled into a latent representation and is input directly into the actor/critic networks, which are trained according to the policy and critic objectives as in standard Soft Actor-Critic [35].

D   Equivariant Adaptation of Policies – Algorithm Pseudocode

---

Algorithm 1 Equivariant Adaptation

Input: Pretrained agent $\mathcal{A}_\theta$ (encoder f, latent model l, policy π), replay buffer $\mathcal{D}$, "oracle" clone of the pretrained agent with frozen weights $\mathcal{A}^*$, canonicalizer C_ϕ i.e. an equivariant CNN, $\mathcal{L}_{prior}$ hyperparameter β, learning rate α, training steps N.

for $i \leftarrow 1$ **to** N **do**
$\quad \mathbf{o} \in \mathbb{R}^{B \times C \times H \times W} \sim \mathcal{D}$
$\quad \mathbf{o}^{canon} \leftarrow C_\phi(\mathbf{o})$
$\quad a \leftarrow \mathcal{A}_\theta(\mathbf{o}^{canon})$
$\quad a^* \leftarrow \mathcal{A}^*(\mathbf{o})$
$\quad z_{cont.}^{canon} \leftarrow f(\mathbf{o}^{canon})$
$\quad z_d^{canon} \leftarrow l(z_{cont.}^{canon})$
$\quad \mathcal{L}_{act} \leftarrow ||a - a^*||_2^2$
$\quad \mathcal{L}_{prior} \leftarrow -D_{KL}(\mathbb{P}_D || P_{C(\mathbf{o})})$
$\quad \mathcal{L}_{commit} = ||\text{StopGrad}(z_d^{canon}) - z_{cont.}^{canon}||_2^2$
$\quad \mathcal{L}_{total} = \mathcal{L}_{act} + \beta \mathcal{L}_{prior} + \mathcal{L}_{commit}$
$\quad \nabla\theta \leftarrow \frac{\partial \mathcal{L}_{total}}{\partial \theta}$
$\quad \theta \leftarrow \theta + \alpha \nabla\theta$
$\quad \nabla\phi \leftarrow \frac{\partial \mathcal{L}_{total}}{\partial \phi}$
$\quad \phi \leftarrow \phi + \alpha \nabla\phi$
end for

Algorithm 1 contains pseudocode for the equivariant adaptation of a pretrained robot policy. A lightweight ENN C_ϕ for the discrete SO(2) symmetry group C_n is initialized and trained to canonicalize the input observation image $\mathbf{o}$, which may be rotated by an arbitrary angle $\frac{360}{n} \cdot i, i \in [0, n]$. The pretrained policy $\mathcal{A}_\theta$ is jointly finetuned to produce optimal actions by minimizing the error between its outputs given the canonicalized observation and the outputs of a cloned version of itself with frozen weights given the original observation.

E   Implementation Details

Parameter	Value
Learning rate	1e-4
Obs horizon	2
Action horizon	8
Prediction horizon	16
Diffusion Embedding Dim	64
Training steps	3e5
Image resolution	64
Number of latents $|z_d|$	20
Values per latent $|V|$	20

Table 3: Task agnostic hyperparameters for ALDA-DP.

Task	# Demonstrations	Episode Length	Camera View
PickCube	1000	100	Angled
PushCube	997	100	Front
PushT	800	200	Front

Table 4: Task specific parameters for ALDA-DP's simulation results.

Task-agnostic and task-specific hyperparameters for ALDA-DP are given in tables 3 and 4, respectively. For a list of ALDA-SAC hyperparameters, we refer the readers to [23]. We use largely the same hyperparameters, with the following exceptions: we do not incorporate framestacking, and the "number of latents" and "values per latent" parameters are set to 10 and 12, respectively.

E.1   Camera View

ManiSkill3 defines the camera location and orientation using the "look at" convention, which accepts as arguments the 3D position of the camera, and the 3D "target" position i.e. where the camera is "looking" with respect to the world frame. For tabletop manipulation tasks, the world frame is roughly the center of the workspace and axis-aligned with the table. By default, these parameters are set to $(0.3, 0.0, 0.6)$ and $(-0.1, 0.0, 0.1)$, i.e. the camera is elevated, front-facing, and pointed downwards roughly towards the center of the workspace. However, this camera view mostly captures the table and very little of the surrounding background, making the "DistractingBackground" visual perturbation less challenging, especially on the PickCube task where we noticed background randomizations having the largest impact on performance. Thus, we define a new "angled" view with the camera and target positions set to $(0.4, 0.5, 0.6)$ and $(0.0, 0.0, 0.35)$, which we use for the PickCube task. This makes the DistractingBackground perturbation more challenging for all methods while maintaining full view of the workspace.

E.2   Real-World Setup

We use a Franka Emika Panda arm, a RealSense D515 camera, and a 3D printed red cube for our main experiments. 200 real demonstrations were collected via teleoperation under visually consistent conditions to train the ALDA-DP model. Proprioceptive state is a 20-dim vector consisting of the arm's joint angles $\in \mathbb{R}^7$, gripper width$\times$2, an "is grasped" boolean $\in \{0, 1\}$, the end-effector pose $\in \mathbb{R}^3$ and quaternion rotation $\in \mathbb{R}^4$, and the goal position $\in \mathbb{R}^3$. Actions are a 4-dimensional vector consisting of the target end effector position $\in \mathbb{R}^3$ and a binary gripper open/close command $\in \{0, 1\}$. This largely mimics the state and action information given by ManiSkill3, except that we do not include joint velocities when training for real-world deployment. All other hyperparameters are the same as the ALDA-DP model trained in simulation.