Learning from 10 Demos: Generalisable and Sample-Efficient Policy Learning with Oriented Affordance Frames

Krishan Rana[†1], **Jad Abou-Chakra**[1], **Sourav Garg**[2], **Robert Lee**, **Ian Reid**[2], **Niko Sünderhauf**[1]
[1]QUT Centre for Robotics, Queensland University of Technology
[2]University of Adelaide
[†]ranak@qut.edu.au

Abstract: Imitation learning has unlocked the potential for robots to exhibit highly dexterous behaviours. However, it still struggles with long-horizon, multi-object tasks due to poor sample efficiency and limited generalisation. Existing methods require a substantial number of demonstrations to cover possible task variations, making them costly and often impractical for real-world deployment. We address this challenge by introducing *oriented affordance frames*, a structured representation for state and action spaces that improves spatial and intra-category generalisation and enables policies to be learned efficiently from only 10 demonstrations. More importantly we show how this abstraction allows for compositional generalisation of independently trained sub-policies to solve long-horizon, multi-object tasks. To seamlessly transition between sub-policies, we introduce the notion of self-progress prediction, which we directly derive from the duration of the training demonstrations. We validate our method across three real-world tasks, each requiring multi-step, multi-object interactions. Despite the small dataset, our policies generalise robustly to unseen object appearances, geometries, and spatial arrangements, achieving high success rates without reliance on exhaustive training data. Video demonstration can be found on our project page: https://affordance-policy.github.io/.

Keywords: behaviour cloning, imitation learning, generalisation, affordances

1 Introduction

Robots operating in domestic environments must solve complex, long-horizon tasks such as preparing a cup of tea, making coffee, or tidying a room, tasks that require coordinating interactions across multiple objects and executing structured action sequences over time. While recent progress in policy learning [1, 2] and large-scale demonstration collection [3, 4, 5, 6] have improved low-level manipulation, imitation learning still struggles with sample efficiency and compositional generalisation. These challenges are amplified in long-horizon, multi-object settings, where task complexity scales quickly and end-to-end policies trained on long, monolithic trajectories often fail to generalise, requiring an impractical number of demonstrations to capture all task variations.

A promising approach to addressing such tasks is to simplify the learning problem towards learning sub-policies that can be composed to solve the longer-horizon task [7, 8, 9, 10, 11]. This however presents several challenges: 1) identifying how to partition the task into sub-tasks that can be independently learned; 2) the distribution shift encountered by sub-policies when presented with the full task configuration and 3) the need to learn an additional arbitration policy that knows when to switch between each sub-policy.

In this work, we take an affordance-centric perspective to address each of these limitations. First, we partition long-horizon, multi-object tasks into affordance-aligned sub-tasks, each defined around a localised object interaction, such as pouring from a teapot or grasping a cup. This provides a

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

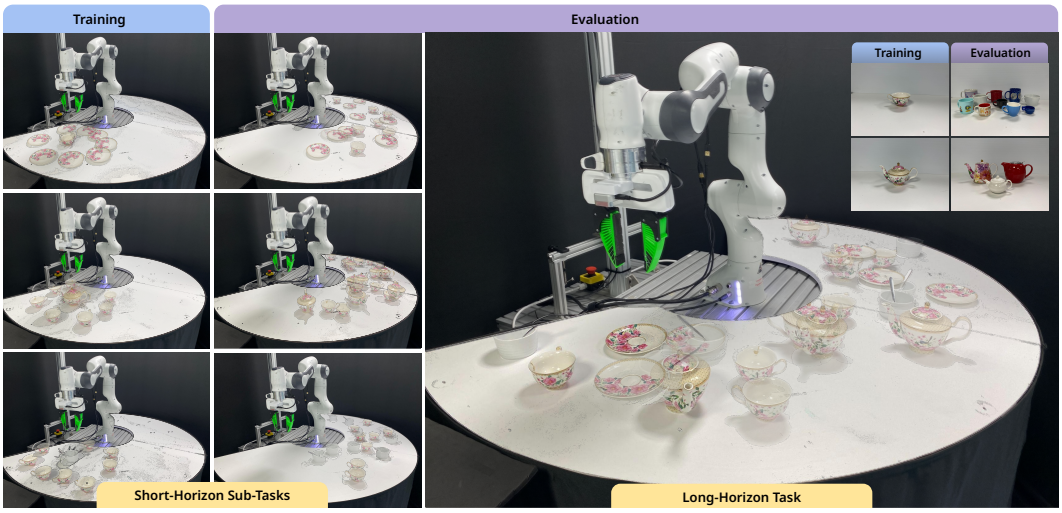


Figure 1: **Compositional Generalisation.** Our representation allows for independent sub-policy training on task-relevant objects from only 10 demonstrations (left). All sub-policies can then be seamlessly composed at evaluation time to solve long-horizon tasks across a vast range of unseen intra-object spatial configurations (right) where all objects are present in the scene as well as intra-category variations from the objects used during training (top-right).

natural and task-relevant decomposition that enables independent training of sub-policies focused on functionally distinct interactions.

Second, we address the distribution shift that arises when composing sub-policies in long-horizon tasks, a challenge exacerbated by two dominant policy representations: image-based inputs [1, 12, 2, 13] and global coordinate frames [13, 12, 14, 15, 16, 17]. Image-based policies trained in isolation fail to capture the full range of visual variations encountered when all task-relevant objects appear at test time, while global frames require demonstrations to cover all possible spatial configurations. To address this, we introduce oriented affordance frames: object-centric coordinate systems anchored at task-relevant affordances and oriented toward the robot's tool frame. These frames retain only the functionally relevant structure of the task, abstracting away clutter and irrelevant details. By rotating the frame with respect to the tool, each sub-policy is trained in a consistent local reference frame, ensuring it remains in-distribution even under novel robot start configurations encountered during policy composition. Grounding policies in these relative frames supports generalisation to spatial variations and novel arrangements without requiring exhaustive demonstration coverage, and naturally enables intra-category generalisation to objects with different appearances or geometries [18, 19, 20, 21]. While affordance representations have been explored in prior work [22, 19, 23, 24], their use in closed-loop behaviour cloning for robust, composable policy learning remains under-explored.

Third, we augment each sub-policy with a continuous self-progress prediction signal, learned directly from the length of demonstration trajectories. This scalar output allows the system to autonomously transition between sub-policies without requiring a separate arbitration policy or external supervision, enabling smooth and robust policy composition across extended task horizons.

Our work makes three key contributions toward scalable and generalisable imitation learning for long-horizon, multi-object manipulation tasks. (1) We introduce the concept of the oriented affordance frame, a local, task-aligned reference frame that enables sub-policy learning to be both spatially invariant and compositionally robust. (2) We develop a perception pipeline that leverages pre-trained vision foundation models to detect and track these affordance frames without reliance on fiducial markers, supporting real-world deployment. (3) We augment each sub-policy with a continuous self-progress prediction signal, enabling automatic and reliable arbitration between sub-tasks without requiring a high-level controller.

Through real-world experiments, we show that our affordance-centric approach enables sample-efficient policy learning from just 10 demonstrations per sub-task, while significantly outperforming image-based and global-frame baselines. It generalises robustly to unseen spatial configurations and novel object instances, and supports seamless composition of independently trained sub-policies.

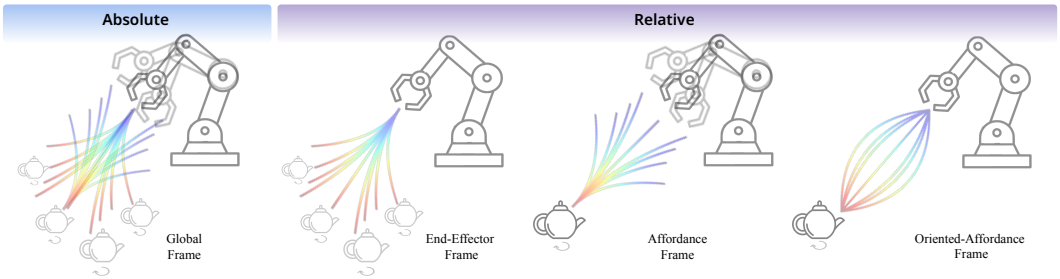


Figure 2: **Comparison of different reference frames for policy learning.** A global reference frame (left) requires demonstrations covering all spatial variations of both the object and the end-effector, leading to poor spatial generalisation. End-effector and affordance-centric frames (middle) reduce this requirement but still require extensive data to capture relative transformations. The proposed oriented affordance frame (right) aligns the state and action representation with the task-relevant affordance and tool frames, thus ensuring spatial invariance while minimising data requirements for policy learning.

Additionally, our marker-free perception pipeline maintains high performance, demonstrating the practicality of our approach in realistic settings.

2 Related Work

Generalisation in Behaviour Cloning: Recent advances in generative modelling have revived interest in behaviour cloning for learning complex, multi-modal behaviours from demonstrations [1, 4, 3, 2]. Behaviour cloning typically maps input states often images or point clouds due to their generality and ease of collection [1, 12, 2, 13] to actions. A major challenge is covariate shift, where small differences between training and test inputs, especially in high-dimensional image spaces, can degrade performance [25, 26]. Current efforts in generalisation focus on large-scale data collection [13, 15, 16, 12] or architectural invariances [27, 28, 29]. However, these approaches mainly tackle spatial generalisation and often retain task-irrelevant details, limiting policy compositionality. We instead propose learning affordance-centric 3D task frames that discard irrelevant information and enable robust intra-category, spatial, and compositional generalisation.

Keypoint-based Representations for Manipulation: Keypoints have been widely used in robotic manipulation to enable intra-category generalisation by focusing on task-relevant object regions [22, 30, 31, 32, 33, 23, 24, 34]. Early approaches trained custom vision models to detect keypoints and solved task-specific SE(3) optimisations for single-step, open-loop tasks [22, 33]. More recent methods [35, 24] leverage pre-trained vision models to extract keypoints or segmentations [36], reducing the need for task-specific training but still operating in open-loop settings with limited spatial invariance. In contrast, we focus on closed-loop behaviour cloning, using keypoint regions as 3D task frames to achieve both spatial and intra-category invariance. We additionally propose a general, task-agnostic pipeline that leverages foundation models to extract keypoint regions without training custom models.

Task Frames: Task frames have long been used in classical robotics to define motions relative to objects or tools to simplify motion generation [37, 38, 39, 40, 41, 42]. Recent works have adapted this idea to reinforcement learning and behaviour cloning. Chi et al. [5] introduced an end-effector–based task frame to simplify in-the-wild data collection, but it still relied on image state representations, which lack task-centric invariances, requiring large-scale demonstrations to generalise. Ke et al. [18] improved spatial invariance by attaching frames to object centres, preserving relative transformations and improving data efficiency for simple tasks. However, their method did not account for object rotations or intra-category variations, limiting generalisation with a tendency to violate robot kinematics when the object rotated beyond certain limits. We address these limitations with a simple approach that supports arbitrary object orientations and generalisation across object instances by introducing an oriented affordance task frame for behaviour cloning.

3 Affordance-Centric Policy Learning with Oriented Affordance Frames

The goal of our work is to train state-conditioned robotic policies $\pi(\mathbf{a}_t|\mathbf{s}_t)$ that are 1) sample-efficient, i.e. they can be learned from as little as 10 human demonstrations; 2) invariant to the

spatial configuration of the task-relevant objects; 3) invariant to variations in the object geometry and appearance; and 4) composable to solve multi-step and long-horizon tasks involving multiple interacting objects.

To achieve this goal, we replace the image- or point cloud- based state representation that is currently prevalent in many imitation learning approaches [1, 12, 2, 13, 43, 44]. Instead, we represent the state $\mathbf{s}_t$ as the pose of the currently task-relevant *tool frame* relative to an *oriented affordance frame*. The latter is a reference frame that is centred on the currently relevant affordance *and* oriented towards the origin of the tool frame at the start of the task. We will describe these coordinate frames in detail in the following (Sec. 3.1), before introducing a perception pipeline that can automatically detect and track these frames on objects unseen during training (Sec. A.1). In Section 3.2, we will describe our proposed method of policy arbitration that enables the autonomous composition of multiple policies to solve long-horizon and multi-step tasks.

3.1 Oriented Affordance Frames for States and Actions

The choice of reference frame for representing state and actions significantly affects a policy's ability to generalise to spatial variations in multi-object tasks, as illustrated in Fig. 2. When states and actions are represented in a fixed global reference frame (first panel), demonstration trajectories must densely cover variations in object and robot poses to attain spatial generalisation. If an object appears in a previously unseen global position, the policy will be out of distribution and likely fail.

Using relative coordinate frames, e.g. expressing the robot's actions relative to its current end-effector or using the pose of task-relevant objects relative to the end-effector, are simple examples of using *relative* reference frames. As illustrated in the middle panels of Fig. 2, these partly alleviate generalisation problems but still require demonstrations to cover all possible poses of task-relevant objects relative to the robot (2nd panel in Fig. 2), or vice-versa (3rd panel), to avoid the policy being out-of-distribution when encountering a previously unseen pose of the object relative to the robot.

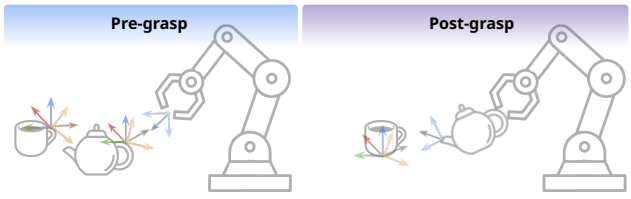


Figure 3: **Affordance Frames, Oriented-Affordance Frames and Tool Frames.** Left: Affordance frames ( ), oriented affordance frames ( ), and tool frame ( ) for a typical *pick* task. Right: Frames for the *pour* task. The oriented affordance frames have the same origin as the affordance frames, but are oriented such that one of the axes (brown) points towards the origin of the tool frame at the beginning of the task.

Affordance Frames: Relative reference frames can not only be centred on the robot's end-effector but also on the affordances of task-relevant objects. Objects can have multiple task-dependent localised *affordances*: e.g. a cup has an affordance on the handle for the task of *picking up* and an affordance in the centre of the cup for the task of *pouring*. Our work makes extensive use of affordance frames, but – importantly and in contrast to previous work [45, 46] – orients them towards the current tool frame and leverage these representations in the behaviour cloning setting.

Tool Frames: In multi-object tasks, a robot either directly interacts with an object (e.g. when picking it up, pushing or opening it), or acts on a target object while holding a *tool* object, e.g. a spoon. In addition to the affordance frame defined on to the *target* object, we define a *tool frame* on the *tool object*. For simple pickup tasks, the tool frame is identical to the robot's end-effector frame, however for actions such as stirring tea with a spoon or pouring from a teapot, the tool frame is placed on the scoop of the spoon or the spout of the teapot respectively as illustrated in Fig. 3.

Oriented-Affordance Frames: Given affordance frames and tool frames, we can now introduce the *oriented-affordance frame*, a core concept of our paper. The oriented affordance frame is obtained by rotating the affordance frame on the target object such that one of its axes (we consistently choose the x-axis as this "funnel" axis) is directed towards the origin of the tool frame.

The oriented affordance frame is the central concept for our generalisable and sample-efficient policy learning: we represent both the state $\mathbf{s}_t$ and the action sequences $\mathbf{a}_t$ of our trained policies $\pi(\mathbf{a}_t|\mathbf{s}_t)$ in the oriented affordance frame.

Frame Initialisation and Update: We initialise the oriented affordance frame at the start of each task. Our perception pipeline (detailed in Appendix A.1) extracts the pose of the currently relevant affordance frame ${}^{\mathrm{W}}\mathbf{T}_{\text{afford}} \in \mathbf{SE}(3)$ in a global world reference frame. With knowledge of the forward kinematics of the robot and the currently held tool object (if any), we also know the pose of the tool frame ${}^{\mathrm{W}}\mathbf{T}_{\text{tool}}$. Using Algorithm 1 we calculate the rotation matrix $\mathbf{R}_{\text{align}}$ that transforms the affordance frame such that its x-axis points towards the origin of the tool frame, thus yielding the oriented affordance frame ${}^{\mathrm{W}}\mathbf{T}_{\text{o-aff}} = \mathbf{R}_{\text{align}} \cdot {}^{\mathrm{W}}\mathbf{T}_{\text{afford}}$.

While the origin of the oriented affordance frame can move during task execution if the robot moves the target object, its *orientation* is kept anchored so that the x-axis keeps pointing to where the origin of the tool frame was *at the beginning of the task*. Our experimental ablation in Sec. 4.2 will demonstrate the benefit of this small but important detail.

State Representation: The state $\mathbf{s}_t$ for our policy comprises the current pose of the tool-frame in the oriented affordance frame ${}^{\text{o-aff}}\mathbf{T}_{\text{tool}} \in \mathbf{SE}(3)$, the binary gripper state $g_{\text{s}} \in \{0, 1\}$, and the rotation ${}^{\text{o-aff}}\mathbf{R}_{\text{aff}}$ of the target object relative to its anchored oriented-affordance frame.

Action Representation: The actions generated by the policy consist of a sequence of $N = 16$ desired next poses of the robot's end effector in the oriented affordance frame $\{{}^{\text{o-aff}}\mathbf{T}_{\text{ee}}\}_{\tau=t\ldots t+N}$ and a sequence of gripper actions $g_{\text{a}} \in \{0, 1\}$ of equal length N.

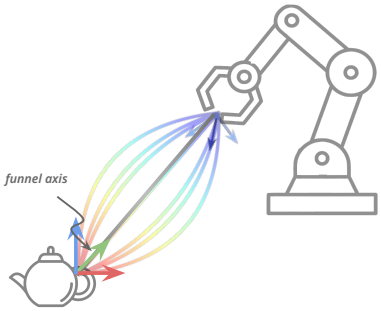


Figure 4: **Adaptive Data Support of the Oriented Affordance Frame.** By aligning the affordance frame with the tool frame at task initiation, demonstrated trajectories become aligned along a consistent 'funnel' axis, reducing variability and improving generalisation. This ensures that policies trained in the oriented affordance frame remain robust to changes in object rotation and robot poses, facilitating seamless policy composition.

Intuitive Benefits of the Oriented Affordance Frame: Expressing robot state and actions relative to the oriented affordance frame maximises the utility of a small number of demonstrations. Intuitively, all demonstrated trajectories tend to be aligned or, to some degree, in the vicinity of the oriented ("funnel") axis, independent of the relative poses of the robot and target object, as illustrated in Figure 4. When composing multiple policies, the oriented affordance frame representation ensures that the robot's tool frame at the end of a task is always *in distribution*, i.e. within the data support of the following policy regardless of the end-effector's global location or the absolute pose of the target object.

3.2 Policy Arbitration by Self-Progress Prediction

With the appropriate abstractions in place, we can now train independent, affordance-centric sub-policies that can be composed to solve longer-horizon tasks. To support automatic policy composition and arbitration, we augment the action space and add a scalar policy *self-progress indicator* $a_{\text{progress}} \in [0, 1]$. During policy training, we compute a task progress measure for a demonstration trajectory by linearly interpolating from 0 to 1 based on the duration of the trajectory. The policy is then trained to output actions and the corresponding progress value in the added progress indicator a_{progress}. During policy execution, this self-progress estimate determines when to transition from one sub-policy to the next, based on a simple threshold. This lets us compose sub-policies to solve complex long-horizon tasks without training an additional arbitration policy.

4 Evaluation

We describe the extensive experiments conducted to support the key claims our paper makes regarding (i) sample-efficient policy training from as little as 10 demonstrations, achieving substantially better performance than other representations; (ii) spatial generalisation; (iii) generalisation to new objects unseen during training; and (iv) the automatic arbitration between sub-policies in long-horizon, multi-step tasks.

We focused our experiments on three multi-step, multi-object tasks representative of scenarios that future domestic service robots are likely to encounter. The first task, preparing a cup of tea, is the primary focus for our quantitative analysis. Additionally, two supplementary tasks – making coffee

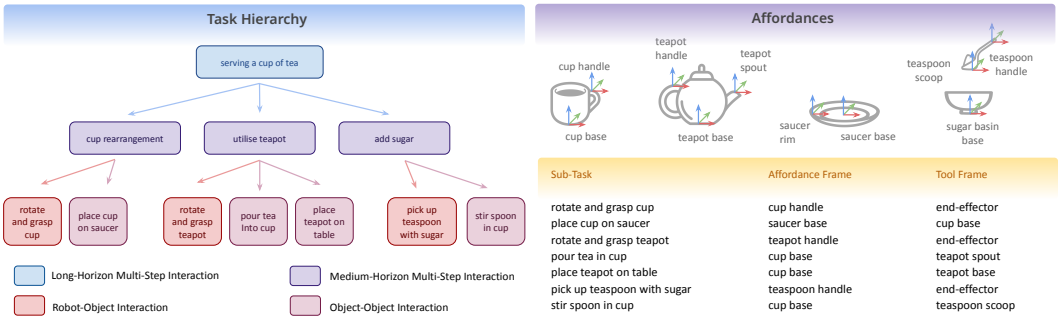


Figure 5: **Affordance-centric task decomposition for the tea serving task.** *Left:* Task decomposition hierarchy; *Top Right:* Affordance-centric frames for each object; *Bottom Right:* Sub-task frame definitions.

and putting a pair of shoes on a rack – are included for qualitative evaluation. Videos demonstrating the trained policies autonomously executing all tasks are available on the anonymous project page: https://affordance-policy.github.io/.

4.1 Experimental Setup – Tea Making

Task Description: We focus our quantitative analysis on a tea-serving task involving five objects: teacup, saucer, teapot, sugar bowl, and teaspoon. The task is decomposed into seven sequential sub-tasks, each defined by the target object and relevant affordance: (1) rotating and grasping the teacup; (2) placing the cup on the saucer; (3) rotating and grasping the teapot; (4) pouring tea into the cup; (5) placing the teapot on the table; (6) scooping sugar with the spoon; and (7) transferring sugar into the cup, stirring, and placing the spoon on the saucer.

We evaluate our method and baselines across these sub-tasks, their compositions (e.g., combining tasks 3–5 into *utilise teapot*), and the full tea-serving sequence, as shown in Fig.5. For each object, we define the corresponding affordance and tool frames (Fig.5). As noted in Appendix A.2, we assume the human demonstrator can identify affordances and meaningful task partitions.

Perception System: We evaluate our method using two perception setups, explicitly indicated for each experiment (Sec.4.2):
1) Marker-based: To isolate the impact of the oriented affordance frames from perception performance (AppendixA.1), we use ArUco markers to obtain ground-truth affordance poses.
2) Large Vision Models: A subset of experiments—including those shown in the supplementary videos—employ our proposed perception pipeline based on pre-trained vision models (Appendix A.1, Fig. 9).

Policy Training: We use Diffusion Policy [1] for imitation learning, training each policy for 4500 epochs with the default parameters from the original implementation. The 16-dimensional state space includes the robot's tool frame pose relative to the oriented affordance frame ${}^{\text{o-aff}}\mathbf{T}_{\text{tool}}$, the binary gripper state $g_s \in 0, 1$, and the object's orientation relative to the oriented affordance frame ${}^{\text{o-aff}}\mathbf{R}_{\text{aff}}$. Both ${}^{\text{o-aff}}\mathbf{R}_{\text{aff}}$ and the rotation component of ${}^{\text{o-aff}}\mathbf{T}_{\text{tool}}$ are expressed as 6D vectors, following [47]. For baselines, we additionally provide the 3D position of the current target object.

The action space is 11-dimensional, comprising the 3D robot position, 6D robot orientation [47], the 1D gripper action, and the 1D self-progress prediction. Our method represents the end-effector pose in the oriented affordance frame, while baselines use either the end-effector frame, the affordance frame or the global frame. Following Diffusion Policy's temporal action generation, the policy outputs a sequence of 16 actions, resulting in a 176-dimensional output vector.

4.2 Results

We report the results of our key experiments and ablation studies below. Each set of experiments supports one of the core claims of our paper.

Sample-efficient Policy Learning from Only 10 Demonstrations: Our first experiment demonstrates that the proposed oriented affordance frame enables highly sample-efficient policy learning from just 10 demonstrations. We evaluated this on seven tasks from the tea-serving scenario, along

Table 1: **Summary of Results.** Success rates for in-distribution (IND) and out-of-distribution (OOD) scenarios for various tasks and composite tasks.

		Oriented Affordance Frame (Ours)		End Effector Frame		Global Frame	
Task	**Demos**	**IND Success**	**OOD Success**	**IND Success**	**OOD Success**	**IND Success**	**OOD Success**
Base Tasks							
(T1) rotate and grasp cup	10	81.8%	81.8%	45.5%	45.5%	45.5%	0.0%
(T2) place cup on saucer	10	100%	100%	100%	100%	9.1%	0.0%
(T3) rotate and grasp teapot	10	90.9%	81.8%	27.3%	27.3%	81.8%	0.0%
(T4) pour tea into cup	10	100%	81.8%	45.5%	27.3%	54.5%	0.0%
(T5) place teapot on table	10	90.9%	72.7%	54.5%	54.5%	90.9%	0.0%
(T6) pick up teaspoon with sugar	10	81.8%	81.8%	45.5%	27.3%	72.7%	0.0%
(T7) stir spoon in cup	10	90.9%	81.8%	18.2%	9.1%	72.7%	0.0%
Average	10	**90.9%**	**83.1%**	48.1%	41.6%	59.7%	0.0%
Composite Tasks							
(T1+T2) cup rearrangement	10	81.8%	81.8%	45.5%	0.0%	36.4%	0.0%
(T3+T4+T5) utilise teapot	10	81.8%	63.6%	18.2%	9.1%	45.5%	0.0%
(T6+T7) add sugar	10	72.2%	72.2%	9.1%	9.1%	63.6%	0.0%
Average	10	**78.8%**	**72.7%**	24.2%	6.1%	48.5%	0.0%
Complete Task							
(T1+T2+T3+T4+T5+T6+T7) serve tea	10	**81.8%**	**63.6%**	9.1%	9.1%	0.0%	0.0%

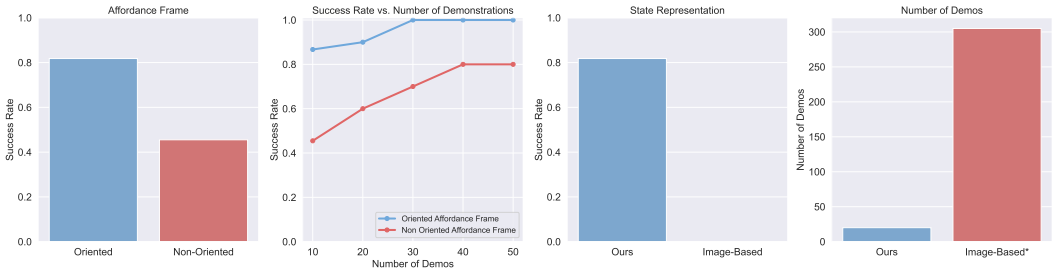


Figure 6: **Additional Comparisons.** *a)* Comparison on different affordance frames; *b)* Success rate vs. number of demonstrations for the different affordance frames; *c)* Performance of an image-based RGB policy when trained with only 10 demonstrations for the cup rearrangement task; *d)* Relative number of demonstrations required for standard image-based diffusion policy [5] to achieve the same generalisation and performance as our system.

with their composite sequences (Fig. 5), using ArUco markers to obtain ground-truth poses for the affordance frames. As shown in Table 1, our method achieved a 90.9% average success rate across all seven tasks, outperforming the end-effector (48.1%) and global (59.7%) frame baselines. This performance advantage persisted across composite tasks, with our method also exceeding 80% success on the full tea-serving sequence, comprising all seven subtasks in order.

We further examined the *cup rearrangement* task, which combines rotating and grasping the cup followed by placing it on the saucer. Fig. 6 shows that an image-based policy similar to [5] failed completely with 10 demonstrations, requiring 305 demonstrations to match our 81.8% success—representing a 30× increase in data requirements. Finally, we ablated the orientation component of our frame representation. Replacing ${}^{\text{o-aff}}\mathbf{T}_{\text{tool}}$ with a non-oriented version (${}^{\text{afford}}\mathbf{T}_{\text{tool}}$) nearly halved success rates at 10 demos and failed to surpass 80% even with 50. In contrast, our method reached 100% success with just 30 demonstrations, highlighting the critical role of the oriented frame in maximising the utility of a small number of demonstrations.

Spatial Generalisation: Our second experiment shows that training with the oriented affordance frame representation leads to better spatial generalisation than other representations, especially when training from a few demonstrations. While the experiment described above evaluated the learned policies under in-distribution conditions (denoted IND in Table 1) where the objects were placed in the same sector of the robot's workspace during training and evaluation, we now vary the spatial configurations and place the objects in different parts of the robot workspace during evaluation. See Fig. 1 for a visualisation. We again use the fiducial markers to provide ground-truth poses of the affordance frames. Under these Out-of-Distribution (OOD) conditions, our proposed representation again performs the best, achieving 83.1% success on average across all base seven tasks, 72.7% on the composite tasks, and still 63.6% on the overall tea-serving task that composes all seven task. The end-effector-centric representation performs much worse (41.6% on average for the base tasks, 6.1% for the compositions and 9.1% for the complete tea-serving scenario) and representing state

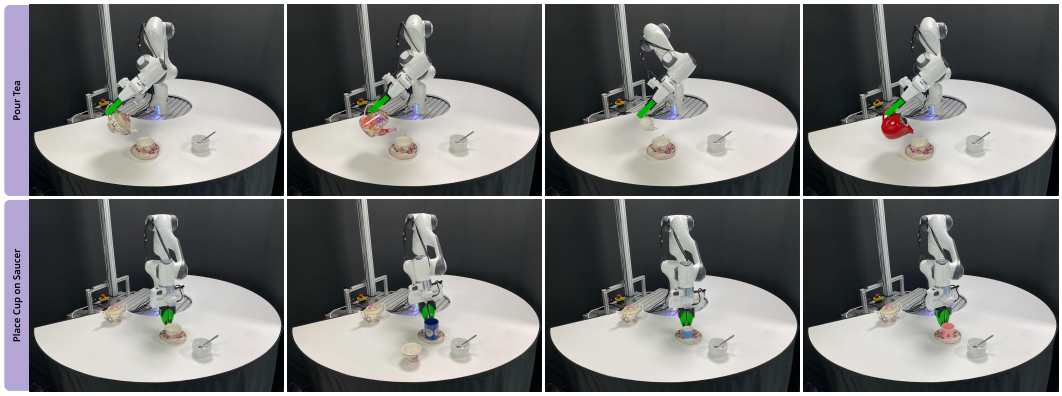
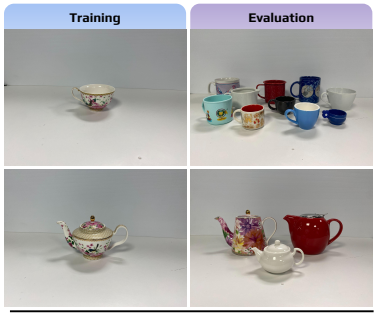


Figure 8: **Intra-category generalisation.** We demonstrate the ability of our approach to enable generalisation across large shape and size intra-category variations.

and action in a global frame fails completely throughout all experiments. A detailed breakdown is provided in Table 1.

Intra-Category Generalisation: Our fourth set of experiments supports our claim that oriented affordance frames enable the transfer of trained policies to new objects unseen during training. These experiments use the perception system from A.1.

As illustrated in Fig. 7, we collected all demonstrations for the *cup rearrange* and *utilise teapot* composite tasks with a single teapot and teacup. The learned policies successfully executed on 8 of the 10 unseen teacups and all 3 unseen teapots. These new objects vary significantly in appearance and geometry compared to the demonstration objects. This successful intra-category generalisation is possible due to the proposed perception pipeline's ability to identify and transfer affordances from one object to another despite considerable intra-category variations in geometry and appearance. See Fig. 8 for further illustration of the generalisation in action.



Task	# Instances	Success
cup rearrange	10	8/10
utilise teapot	3	3/3

Figure 7: **Generalisation to intra-category variations** The set of objects used for training and evaluating the intra-category generalisation capabilities of the trained sub-policies.

Task Composition and Self-Progress Prediction: Throughout our experiments involving composite tasks (e.g. *cup rearrangement* or the full tea-making task), we utilise the proposed self-progress prediction mechanism to fully autonomously control the transition between base tasks. We found the self-progress prediction to be remarkably robust in all experiments and did not observe it to cause any failures. More detailed results are provided in Appendix A.7.

5 Conclusion

Our experiments demonstrate that oriented affordance frames substantially enhance both sample efficiency and generalisation in imitation learning for long-horizon, multi-object tasks. By replacing dense image- or point-cloud-based representations with an abstracted, affordance-centric formulation, our approach enables robust policy learning from as few as 10 demonstrations. It generalises effectively across spatial, intra-category, and combinatorial variations - crucial for solving complex, long-horizon tasks. We also introduce a perception pipeline to detect and track these frames using vision foundation models, along with a simple yet effective progress estimation metric derived from demonstration duration to enable seamless sub-policy transitions. We hope this work encourages further exploration of structured representations, priors, and compositionality in behaviour cloning, paving the way toward more generalisable and practical robotic systems for real-world applications.

Acknowledgments

The authors also acknowledge the ongoing support from the QUT Centre for Robotics. This work was partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (Project DP220102398) and by an Amazon Research Award to Niko Sünderhauf. This work was also supported by the QUT Research Engineering Facility.

References

[1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[2] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi:10.15607/RSS.2023.XIX.016.

[4] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *arXiv*, 2024.

[5] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[6] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators, 2023.

[7] Z. Zhang, Y. Li, O. Bastani, A. Gupta, D. Jayaraman, Y. J. Ma, and L. Weihs. Universal visual decomposer: Long-horizon manipulation made easy. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6973–6980. IEEE, 2024.

[8] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

[9] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *The International Journal of Robotics Research*, page 02783649241304789, 2023.

[10] Y. Lee, E. S. Hu, and J. J. Lim. Ikea furniture assembly environment for long-horizon complex manipulation tasks. In *2021 ieee international conference on robotics and automation (icra)*, pages 6343–6349. IEEE, 2021.

[11] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations. *arXiv preprint arXiv:2003.06085*, 2020.

[12] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[13] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[14] F. Ceola. Robotic perception and manipulation: Leveraging deep learning methods for efficient instance segmentation and multi-fingered grasping. 2024.

[15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[16] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[17] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

[18] L. Ke, J. Wang, T. Bhattacharjee, B. Boots, and S. Srinivasa. Grasping with chopsticks: Combating covariate shift in model-free imitation learning for fine manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6185–6191. IEEE, 2021.

[19] N. D. Palo and E. Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[20] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=9iG3SEbMnL.

[21] M. Pan, J. Zhang, T. Wu, Y. Zhao, W. Gao, and H. Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. *arXiv preprint arXiv:2501.03841*, 2025.

[22] L. Manuelli, W. Gao, P. Florence, and R. Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019.

[23] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann. Neural descriptor fields: Se(3)-equivariant object representations for manipulation. 2022.

[24] Y. Wang, M. Zhang, Z. Li, T. Kelestemur, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li. D^3fields: Dynamic 3d descriptor fields for zero-shot generalizable rearrangement. *Conference on Robot Learning (CoRL)*, 2024.

[25] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[26] M. Torne, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024.

[27] M. Weiler and G. Cesa. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

[28] D. Wang, R. Walters, X. Zhu, and R. Platt. Equivariant q learning in spatial action spaces. In *Conference on Robot Learning*, pages 1713–1723. PMLR, 2022.

[29] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt. Sample efficient grasp learning using equivariant models. *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[30] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Conference on Robot Learning*, pages 1602–1611. PMLR, 2021.

[31] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *arXiv preprint arXiv:2104.01542*, 2021.

[32] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih. Unsupervised learning of object keypoints for perception and control. *Advances in neural information processing systems*, 32, 2019.

[33] P. R. Florence, L. Manuelli, and R. Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv preprint arXiv:1806.08756*, 2018.

[34] M. Sharma and O. Kroemer. Generalizing object-centric task-axes controllers using keypoints. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7548–7554, 2021. doi:10.1109/ICRA48506.2021.9561577.

[35] N. Di Palo and E. Johns. Keypoint action tokens enable in-context imitation learning in robotics. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[36] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Object-centric imitation learning for vision-based robot manipulation. In *6th Annual Conference on Robot Learning*, 2022.

[37] D. H. Ballard. Task frames in robot manipulation. In *AAAI*, volume 19, page 109, 1984.

[38] M. H. Raibert and J. J. Craig. Hybrid position/force control of manipulators. 1981.

[39] M. T. Mason. Compliance and force control for computer controlled manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(6):418–432, 1981.

[40] D. Berenson, S. Srinivasa, and J. Kuffner. Task space regions: A framework for pose-constrained manipulation planning. *The International Journal of Robotics Research*, 30(12): 1435–1460, 2011.

[41] J. E. King, M. Cognetti, and S. S. Srinivasa. Rearrangement planning using object-centric and robot-centric action spaces. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3940–3947. IEEE, 2016.

[42] T. Migimatsu and J. Bohg. Object-centric task and motion planning in dynamic environments. *IEEE Robotics and Automation Letters*, 5(2):844–851, 2020.

[43] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *Arxiv*, 2024.

[44] 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 *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[45] J. Gao, Z. Tao, N. Jaquier, and T. Asfour. K-vil: Keypoints-based visual imitation learning. *IEEE Transactions on Robotics*, 2023.

[46] W. Gao and R. Tedrake. kpam 2.0: Feedback control for category-level robotic manipulation. *IEEE Robotics and Automation Letters*, 6(2):2962–2969, 2021.

[47] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

[48] J. K. S. B. Bowen Wen, Wei Yang. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.

[49] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[50] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[51] CSMCube. Csm cube. URL `https://www.csm.ai/`.

[52] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep ViT Features as Dense Visual Descriptors. *European Conference of Computer Vision Workshop (ECCVW) on What is Motion For?*, 2022.

[53] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[54] J. Gao, X. Jin, F. Krebs, N. Jaquier, and T. Asfour. Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16850–16857. IEEE, 2024.

[55] T. W. T. B. A. V. Nick Heppert, Max Argus. Ditto: Demonstration imitation by trajectory transformation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

A Appendix

A.1 A Perception Pipeline to Detect and Track Affordance Frames

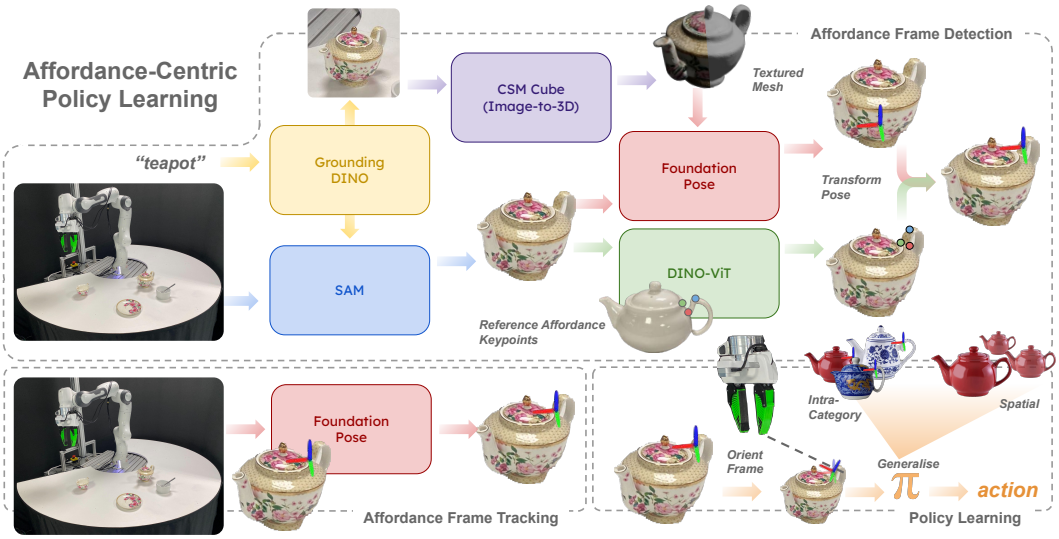


Figure 9: **Affordance-Centric Policy Learning.** *Affordance Frame Detection:* We propose a framework to detect affordance frames using pre-trained large vision models. *Affordance Tracking:* Once the frame is detected we utilise Foundation Pose [48] to continuously track the frame in real-time as the robot interacts with it. *Policy Learning:* At the start of each episode we appropriately orient the frame towards the tool frame of the robot and train a state-based diffusion policy that operates with this frame as its task frame.

Automatically detecting and tracking the affordance frames necessary for the affordance-centric policy learning presented in the previous section is not a trivial task. However, recent progress in computer vision and especially in generalist vision foundation models, makes this possible now. In this section, we present a perception pipeline that can detect and track affordance frames.

We note that we specifically do not claim the following approach to be superior to alternative methods. We do not present an in-depth analysis or comparison to alternatives, as this is beyond the scope of our paper. We found that the pipeline presented in the following was effective in our setup, but future work definitely can improve upon what we present here. A complete visual overview of our perception pipeline is given in Fig. 9.

Assumptions: We assume that for each task, the human demonstrator provides the following input during the policy training stage to aid the perception system later during autonomous policy execution: (1) the name of the target object the robot has to interact with, e.g. "*teapot*"; and (2) three reference points that localise the affordance-frame of the task.

Affordance Frame Detection: At the start of each subtask, given an input image of the robot's workspace from an external camera (see Fig. 13) and the user-provided name of the target object, we first use Grounding DINO [49] to detect a bounding box around the object, leveraging its open-vocabulary capabilities. The image is then cropped to this bounding box and passed to SAM [50] to obtain a segmentation mask of the object. We additionally pass the cropped image to CSM Cube's [51] Image-to-3D model, which generates a textured mesh of the object. Both the segmentation mask and the textured mesh are then used to initialise Foundation Pose [48] for object pose estimation.

However, this pose is extracted relative to the mesh centre rather than the specific affordance relevant to the task. To localise the affordance pose, we extract DINO-ViT features [52] from the current image crop and match them against the DINO-ViT features of the three reference points provided by the human demonstrator during policy training. This matching process allows us to transform the mesh-centred pose detected by Foundation Pose to align with the affordance region. This process is illustrated in Fig. 9 (top).

	Success Rate	Type of Error		
		Joint Limit Violation	Out of Distribution	Tracking Error
Affordance Frames				
Oriented	**82%**	0.0%	100%	0.0%
Non-Oriented	**46%**	36.4%	63.6%	0.0%
Perception System				
Aruco Markers	**80%**	0.0%	100%	0.0%
Foundation Pose [48]	**70%**	0.0%	66.7%	33.3%

Table 2: **Ablation Study.** Analysis of failure modes when comparing the two different affordance frames and perception systems.

This entire process is performed once at the start of the task. For long-horizon tasks involving multiple objects and sequential subtasks, we initialise all affordances at the beginning and track them in parallel using multiple Foundation Pose models.

Affordance Frame Tracking: Once initialised, we continuously track the pose of an object using Foundation Pose [48] and transform it to the affordance region at approximately 20 Hz on a desktop computer with an RTX4090 GPU. This tracked frame is used to compute the oriented affordance frame in which states and actions are represented. As we switch between subtasks for long-horizon tasks, we transition between the already-initialised affordance frames required for the respective subtasks.

Efficacy of Perception Pipeline: In the experiments, we decoupled the effects of the perception pipeline (described in Appendix A.1) on the policy performance and used fiducial markers on the objects to mark the pose of the various affordance frames. In this third set of experiments (Table 2), we show that our proposed perception pipeline is able to detect and track affordance frames without the use of fiducial markers, with only minimal decrease in task performance.

We ran 10 trials of the *cup rearrange* task, which is a composition of the base tasks of rotating and grasping the tea cup and placing it on the saucer. When using the fiducial markers, 8 out of 10 trials were successful, which is consistent with the results reported in Table 1, as expected. Removing the fiducial markers and using the proposed perception pipeline to detect and track the affordance frames resulted in 7 successful trials, indicating that only one additional failure case was introduced by the full perception pipeline. This failure case was a tracking error, where the vision system lost track of the objects pose due to occlusions from the robot, whereas the other two failure cases were a result of the policy failing to successfully grasp the cup or stagnation typically seen when the policy falls out of distribution. Further qualitative results in the accompanying videos show that we can train policies from 10 demonstrations and successfully execute them without any fiducial markers for the full tea-serving task, as well as the shoe-racking and coffee-making tasks, as illustrated in Fig. 12.

A.2 Assumptions and Limitations

There is no free lunch [53], and the reduction in required demonstrations while gaining spatial and intra-category invariance does not come for free. While our proposed approach significantly reduces the burden on the human demonstrator to provide a large number of task demonstrations, we make the following assumptions:

(1) The objects involved in the tasks have clearly defined affordances, and a human demonstrator can identify the location of the relevant affordance frames. We find this to be a mild assumption for many objects involved in typical tasks in a domestic scenario, but we acknowledge that some tasks (e.g. laundry folding) will break this assumption.

(2) The affordance frames appear at locations on the object that are distinct and informative enough for a perception pipeline to identify and track them, as well as transfer them across objects within the same category. We found this assumption to hold well for the evaluated real-world tasks, but objects without characteristic geometries or appearance will pose a challenge.

(3) The human demonstrator can identify sub-tasks within the long-horizon task. This is a very mild assumption, and the partitioning of tasks could be automated based on detecting when the robot starts interacting with the next object, or even with the help of Large Language Models.

Limitations: While our evaluation showed the proposed approach to be effective in learning long-horizon tasks, there are some noteworthy limitations that warrant further exploration. Most importantly, the proposed method depends on reliable object tracking to continually update the pose of the affordance frame during a manipulation task. While the presented perception pipeline from Sec. A.1 worked well in the tested scenarios, it has limitations when tracking through occlusions, or dealing with non-rigid (e.g. articulated or deformable) objects.

Second, the pose-based abstraction of objects could limit the applicability to tasks not easily represented by object affordance frames alone, potentially requiring additional modalities like tactile sensing to capture more fine-grained object details. Despite these limitations, our contribution offers a promising pathway to more sample-efficient and generalisable imitation learning of complex long-horizon manipulation tasks.

Informally, our approach relieves the human demonstrator from the burden of collecting a large number of diverse demonstrations and reduces the pressure on the policy learning algorithm to extract task-relevant generalisation information from raw image-based state inputs. Instead, we shift part of the inherent difficulty of imitation learning to a dedicated perception system that can extract and track affordance frames. One might argue that this merely redistributes the difficulty rather than reducing it. However, we are confident that this shift is highly beneficial: acquiring large-scale training data for generalist vision models, such as those used in Sec. A.1, is significantly cheaper and more scalable than collecting extensive human demonstrations and robot interaction data, which remain expensive and labour-intensive. Thus, we expect the performance of specialised vision perception systems to continue to improve rapidly, becoming even more useful in the context of imitation learning soon.

A.3 Applicability to Mobile Manipulation

By training our policy with respect to a relative frame attached to an object, the robot's action and state space remain consistent regardless of the position of the robot's base. This allows for the policy to continue operation while the base of the robot is in motion. We demonstrate this by running the same policy trained in the tabletop setting on a mobile manipulator robot and show how the end effector of the robot can maintain task performance regardless of the movement of the robot's base as illustrated by the discrepancy between the green and red robot base locations in Figure 10. A video of this experiment is provided in the supplementary material.

A.4 Closed Loop Control

Prior work have introduced the concept of local keypoints or regions on objects as compact representations for manipulation tasks [23, 22, 35]. These systems have traditionally been used to define start and end poses for simple pick-and-place operations, utilizing off-the-shelf inverse kinematics and motion planners to move objects from one location to another in an open-loop manner [23, 22]. Other methods have incorporated these representations within the context of imitation learning, primarily focusing on one-shot imitation learning [45, 54, 55]. In these cases, the keypoint locations are used to define complex admittance controllers [45] or prompt large language model (LLM) [35] to replicate a single trajectory, limiting their ability to react to changes or perturbations during policy execution. Our approach in contrast, leverages these representations in a behaviour cloning setting where we can learn closed-loop diffusion policies [1] that are robust to perturbations and allow us to move beyond simple pick-and-place tasks to imitating more complex closed-loop tasks, including non-prehensile manipulation, such as pushing objects as shown in Figure 11 below.

A.5 Tasks

A.6 Experimental Setup – Further Details

For all evaluations, we used a Franka Panda manipulator arm equipped with Intel Realsense cameras as shown in Figure 13. We utilised Cartesian impedance control to control the robot. All demonstrations were collected using a GELLO teleoperation device [6]. During data collection, all

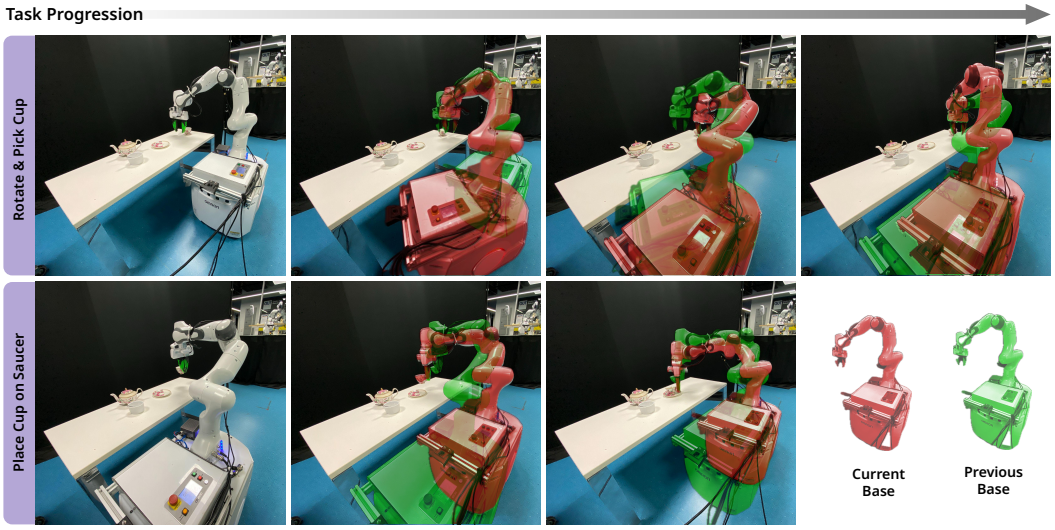


Figure 10: **Robustness to moving base.** We demonstrate our ability to maintain task performance regardless of the robot's moving base when operating with respect to an affordance-centric task frame.

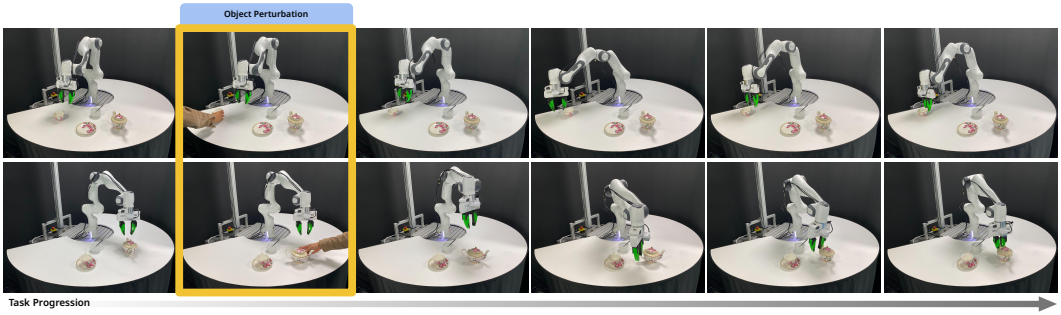


Figure 11: **Closed-loop control.** We demonstrate our ability to learn robust policies that react to object perturbations during execution beyond simple pick-and-place tasks.

objects are equipped with an individual ArUco marker (Figure 13) which we use for identifying the affordance-centric frames via measured rigid transforms from this marker, as well as for tracking these frames across the demonstration. We chose this method to obtain the affordance frame as it allowed us to decouple the performance of the perception system from the utility of the affordance frames for policy learning and composition, which was the main focus of this work. For all the video demonstrations in the supplementary, we switched to the marker-free setup as shown in Figure 13 (right).

A.7 Policy Composition – Further Details

Having trained each affordance-centric policy, we can compose them to solve long-horizon, multi-object tasks. We first define the order of sub-tasks and their associated affordance frames required to complete the full task. The robot then performs an initialisation scan of the environment to identify the initial locations of all objects and their local affordance frames in the scene. Once identified it runs the first policy corresponding to the first sub-task. As the policy is trained to output end-effector poses defined in the oriented affordance frame ${}^{\text{o-aff}}\mathbf{T}_{\text{ee}}$, we transform these actions to the base frame of the robot $\mathbf{T}_{\text{ee}}$ before executing them with a Cartesian impedance controller. If a_{progress} generated by the policy increases beyond a predefined threshold ϕ, indicating sub-task completion, we switch affordance frames and repeat the process with the next policy corresponding to the next sub-task.

Tea Serving Shoe Racking Coffee Making

Figure 12: **Demonstrating our system across three diverse real-world tasks.** The Tea Serving and Coffee Making tasks are very complex compared to many tasks typically encountered in the literature and require multiple sequential object interactions and a high level of interaction precision, e.g. when operating the coffee machine or pouring tea into the cup. Videos of the robot autonomously executing learned policies for all tasks are provided in the supplementary material.

Figure 13: **Experimental Setup.** *Left:* Marker-based setup using 2 wrist-mounted D405 Intel Realsense cameras for top down detection. *Right:* Marker-free setup using a front-facing D455 Intel Realsense camera running Foundation Pose.

Figure 15 illustrates the predicted progress while executing each base task in the tea-serving scenario. The system switches to the next policy when the predicted progress reaches a predefined threshold.

We further tested the responsiveness of this self-progress indicator to external disturbances for the task of rotating a cup and then grasping it. As illustrated in Fig. 14, a human interfered during task execution by moving or rotating the cup. The self-progress prediction value immediately decreases as the task is partially reset and gradually increases as the robot proceeds with solving the task.

A.8 Diffusion Policy

Throughout this work, we leverage diffusion policies [1] as our central behaviour cloning algorithm. Diffusion policy models the conditional action distribution as a denoising diffusion probabilistic model (DDPM), allowing for better representation of the multi-modality in human-collected demonstrations. Specifically, diffusion policy uses DDPM to model the action sequence $p(\mathbf{A}_t \mid \mathbf{o}_t, \mathbf{x}_t)$, where $\mathbf{A}_t = \{\mathbf{a}_t, \ldots, \mathbf{a}_{t+C}\}$ represents a chunk of next C actions. The final action is output of the following denoising process:

$$\mathbf{A}_t^{k-1} = \alpha(\mathbf{A}_t^k - \gamma\epsilon_\theta(\mathbf{o}_t, \mathbf{x}_t, \mathbf{A}_t^k)) + \mathcal{N}(0, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{A}_t^k$ is the denoised action sequence at time k. Denoising starts from $\mathbf{A}_t^K$ sampled from Gaussian noise and is repeated till $k = 1$. In Equation (1), (α, γ, σ) are the parameters of the denoising process and ϵ_θ is the score function trained using the MSE loss $\ell(\theta) = (\epsilon_k - \epsilon_\theta(\mathbf{o}_t, \mathbf{x}_t, \mathbf{A}_t^k + \epsilon_k))^2$. The noise at step k of the diffusion process, ϵ_k, is sampled from a Gaussian of appropriate variance.

Figure 14: **Self-Progress Behaviour with Disturbances.** Predicted progress over the course of a task exposed to two different disturbances and resets mid-execution. The task requires the robot to rotate a coffee mug and then grasp it. We indicate the start of the two disturbances where we reset this rotation by the grey dotted vertical lines in the plot.

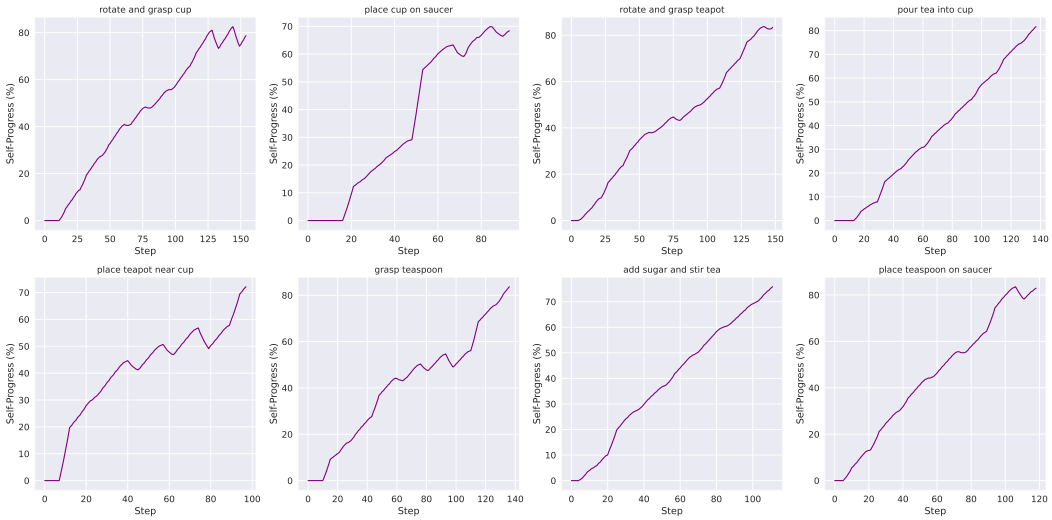


Figure 15: **Self-Progress Predictions across the Tea Making Task.** Each sub-plot indicates the predicted self-progress for a different sub-task in the tea-making task.

The policy predicts a sequence of 16 actions, of which we execute the first 8. The diffusion network has 8.08 million parameters, we used a learning rate of 10^{-4}. The rest of the implementation is identical to the original implementation [1].

Algorithm 1: Calculation of $\mathbf{R}_{\text{align}}$

Input: $\mathbf{p}_{\text{tool}}, \mathbf{p}_{\text{afford}}$
Output: $\mathbf{R}_{\text{align}}$
Function `ComputeRotationMatrix`$(\mathbf{p}_{tool}, \mathbf{p}_{afford})$**:**
  Define the Vectors:
  $\mathbf{v}_{\text{funnel}} \leftarrow [1, 0, 0]^T$
  $\mathbf{p}_{\text{tool}} \leftarrow$ Position of the tool frame
  $\mathbf{p}_{\text{afford}} \leftarrow$ Position of the affordance frame
  Calculate the Direction Vector:
  $\mathbf{d} \leftarrow \mathbf{p}_{\text{tool}} - \mathbf{p}_{\text{afford}}$
  $\mathbf{d}_{\text{norm}} \leftarrow \frac{\mathbf{d}}{\|\mathbf{d}\|}$
  Find the Rotation Axis and Angle:
  $\mathbf{r} \leftarrow \mathbf{v}_{\text{funnel}} \times \mathbf{d}_{\text{norm}}$
  $\mathbf{r}_{\text{norm}} \leftarrow \frac{\mathbf{r}}{\|\mathbf{r}\|}$
  $\cos(\theta) \leftarrow \mathbf{v}_{\text{funnel}} \cdot \mathbf{d}_{\text{norm}}$
  $\sin(\theta) \leftarrow \|\mathbf{r}\|$
  Construct the Rotation Matrix:
  $\mathbf{K} \leftarrow \begin{bmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{bmatrix}$
  $\mathbf{R}_{\text{align}} \leftarrow I + \sin(\theta)\mathbf{K} + (1 - \cos(\theta))\mathbf{K}^2$
  return $\mathbf{R}_{align}$