# Affordance-Centric Policy Learning:
## Generalisable Robot Policy Learning using Affordance-Centric Task Frames

**Anonymous Author(s)**
Affiliation
Address
email

**Abstract:**
Affordances are central to robotic manipulation, where most tasks can be simplified to interactions with task-specific regions on objects. By focusing on these key regions, we can abstract away task-irrelevant information, simplifying the learning process, and enhancing generalisation. In this paper, we propose an affordance-centric policy-learning approach that centres and appropriately *orients* a *task frame* on these affordance regions allowing us to achieve both **intra-category invariance** – where policies can generalise across different instances within the same object category – and **spatial invariance** — which enables consistent performance regardless of object placement in the environment. We propose a method to leverage existing generalist large vision models to extract and track these affordance frames, and demonstrate that our approach can learn manipulation tasks using behaviour cloning from as little as 10 demonstrations, with equivalent generalisation to an image-based policy trained on 305 demonstrations. We provide video demonstrations on our project site: redacted for double-blind.

**Keywords:** Behaviour Cloning, Affordances, Object-Centric Representations, Generalisation, Sample Efficiency

## 1 Introduction

Vision-based robotic manipulation is essential for enabling autonomous robots to operate effectively in unstructured, everyday environments. These environments present numerous challenges, particularly when tasks involve interacting with objects that vary in spatial positioning and exhibit intra-category differences in visual appearance, shape, and size. Recent advances in behaviour cloning [1, 2, 3, 4] have shown promise in enabling robots to learn complex visuomotor policies in these settings by directly mapping raw visual inputs to motor actions without the need for manual feature engineering. However, such end-to-end approaches are highly sensitive to covariate shifts [5, 6, 7, 8, 9] and often overfit to task-irrelevant information present in the images, such as specific visual appearances, object locations, or environmental distractors. This overfitting leads to poor generalisation when the robot encounters new task settings or variations not present in the training data. Consequently, current efforts to mitigate these challenges have focused on collecting large-scale datasets [10, 11, 12, 13, 14, 15] that aim to capture all potential variations, which is both resource-intensive and impractical.

In this work, we seek to address the generalisation challenge for robotic manipulation without the need for large, exhaustive datasets. More specifically, we propose an approach to address both the *spatial* and *intra-category* object generalisation capabilities of behaviour cloning. We build on recent advances in representation learning for robotic manipulation which have shown that local, affordance-centric keypoints can yield significant intra-category generalisation across a wide range of open-loop manipulation tasks [16, 17, 18, 19, 20], and propose a novel closed-loop behaviour cloning formulation using this representation that simultaneously accounts for spatial generalisation. We introduce two key ideas to achieve this: *1)* as opposed to directly feeding this representation as input to the policy [21, 22], we redefine the task frame for policy learning based on a localised coordinate system centred on these affordance-centric regions and *2)* as our policy now operates in

Submitted to the 8th Conference on Robot Learning (CoRL 2024) Workshop on Learning Robotic Assembly of Industrial and Everyday Objects.

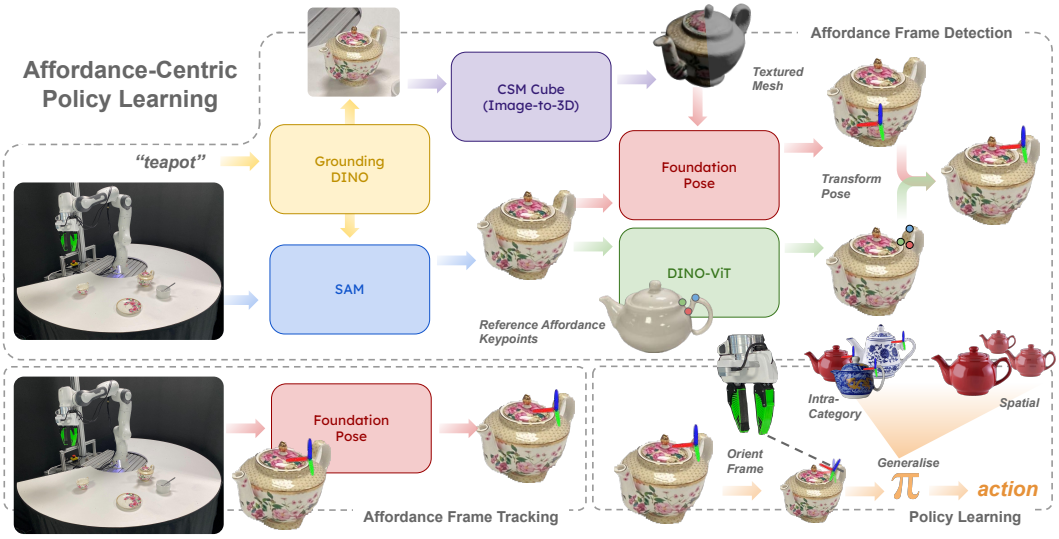



Figure 1: **Affordance-Centric Policy Learning.** *Affordance Detection:* We propose a framework to detect affordance frames using pre-trained large vision models. *Affordance Tracking:* Once the frame is detected we utilise Foundation Pose to continuously track the frame in real-time as the robot interacts 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.

a relative action frame, we anchor the orientation of this frame towards the end-effector of the robot to avoid joint limit violations as the object rotates.

By capturing the spatial location of an object using the task frame, we simplify the state representation needed for policy learning to the $\mathbf{SE}(3)$ pose of the end effector relative to this coordinate frame. This reduces the dimensionality of the search space for policy learning while simplifying data collection where initial states can be as straightforward as rotating an object in-place and recording a demonstration. For most manipulation tasks involving object rearrangement, this simple abstraction is sufficient to learn complex policies without relying on image inputs, significantly reducing and simplifying the data requirements for achieving open-world generalisation.

We introduce a complete perception pipeline to detect and track these affordance-centric frames using pre-trained large vision models and demonstrate our ability to learn a wide range of manipulation tasks from only 10 demonstrations with the ability to generalise to both spatial and intra-category variations.

## 2 Related Work

**Generalisation in Behaviour Cloning:** There has been a recent resurgence in behaviour cloning methods given advances in generative modelling and the ability to capture complex multi-modal behaviours from demonstrations [1, 2, 3, 4]. Behaviour cloning typically learns to map input states either directly or indirectly to actions with images becoming the most ubiquitous state representation given the simplicity and generality they provide for open-world behaviour learning. The key challenge faced by these systems is covariate shift as the test distribution varies from the data that the policy was trained on [5, 6, 7, 8, 9] . Small changes in the input state representation can be detrimental to the performance of the trained policy, which is further exacerbated by the high-dimensional and potential variations that could be exhibited in image-based representations [23]. The ability to generalise to these variations is currently being extensively explored by the robot learning community with most efforts emphasising scaling data collection [10, 11, 12, 13, 14] to capture all these potential variations given the recent successes of large scale training demonstrated by the vision [24, 25, 26, 27] and natural language processing communities [28, 29]. Other works additionally explore inducing these invariance or equivariances directly via the network architecture [30, 31, 32, 33] or via 2D image augmentations [34, 35, 36]. These methods however typically only address the spatial generalisation of policy learning and are limited to 2D space. We focus on abstracting images

to local affordance-centric task frames that naturally operate in 3D space while abstracting away a significant amount of non-task relevant information allowing for both intra-category and spatial generalisation.

**Representations for Manipulation:** Keypoints have been widely used in robotic manipulation to enable intra-category generalisation by focusing on local task-relevant object regions [16, 37, 38, 39, 40, 17, 20, 18]. Prior works train custom models to detect keypoints [16, 40] and solve task-specific optimisation problems in $\mathbf{SE}(3)$ for a wide range of *single-step* manipulation tasks. These methods however require task-specific vision systems trained to obtain the required keypoints and have only been demonstrated to operate in an open-loop control setting. Recent methods [22, 20] have introduced the idea of using pre-trained large-vision models to extract these keypoints without the need for training task-specific systems, however still focus on open-loop control settings or utilise these representations or object-centric segmentations [21] directly as input to a policy. While yielding better intra-category generalisation and sample efficiency, these methods still exhibit poor spatial invariance. In this work, we focus on the closed-loop behaviour cloning setting and leverage these keypoint regions as a task frame for policy learning to attain both spatial and intra-category invariance. Furthermore, we introduce a general task-agnostic pipeline that leverages existing foundation models to obtain these keypoint regions without training custom, single-task models.

**Task Frames:** Task frames have been extensively used in classical robotics to enable robust task execution by defining motions relative to frames attached to objects or tools [41, 42, 43, 44, 45, 46]. Recent works have extended these ideas to reinforcement learning and behaviour cloning. Chi et al. [47] proposed a novel task frame based on the end-effector position to facilitate in-the-wild data collection without relying on a fixed global frame. While this approach simplified data collection, it did not induce task-centric invariances to enhance sample efficiency, still requiring extensive demonstrations across a wide range of spatial and intra-category variations. Ke et al. [48] improved upon this by attaching a relative frame to the centre of an object, transforming the data into an object-centric frame that preserves the relative transformation between the end-effector and object. This made training data denser around critical regions for task success, enhancing spatial invariance in simple pick-and-place tasks. However, they did not consider object rotation or tasks where object instances could vary. This introduces a set of additional challenges where the relative task frame can vary in ways that could violate the kinematic constraints of the underlying robot. We propose a simple approach to address these challenges enabling us to operate across all object orientations while generalising to variations in object instances, allowing for open-world policy generalisation.

# 3 Problem Formulation

We focus on behaviour cloning for robotic manipulation tasks, given access to a set of $N$ demonstrations $D = \tau_i{}_{i=1}^{N}$, where each trajectory $\tau_i$ is demonstrated through teleoperation. The objective is to learn a policy that effectively mimics these actions with the ability to generalise to unseen settings. Achieving generalisation in behaviour cloning is challenging due to the variability in unstructured environments. Specifically, our goal is to address two key challenges: **spatial generalisation**, which involves handling different object placements, and **intra-category generalisation**, which deals with variations across different instances of the same object category. While relative task frames have shown promise for spatial generalisation, and local key points have been successful for intra-category generalisation, we seek a unified framework for policy learning that combines these two ideas while addressing their respective limitations. Key questions we seek to answer include: **1)** How can we obtain the required relative task-frame frames without relying on task-specific models for each scenario? **2)** How can we ensure that the relative task frame supports both spatial and intra-category generalisation? **3)** How can we ensure robust operation across all possible $\mathbf{SO}(3)$ object rotations to which the task frame is attached?

# 4 Affordance-Centric Policy Learning

We propose a unified framework for generalisable policy learning using relative task frames. More specifically, we centre our relative task frame at local affordance-centric regions on objects. The reasons for this are two-fold: 1) These affordance regions holistically capture the interaction points

for a wide range of manipulation tasks allowing us to fully define the state of an object based on the **SE**(3) pose of the frame and 2) this region is invariant across object instance variations including visual appearance, shape and size, allowing for intra-category behaviour generalisation. We describe a general pipeline to obtain these frames and detail how we utilise this frame for policy learning in the following sections.

### 4.1 Affordance Frame Detection and Tracking

We leverage the impressive number of generalist vision foundation models that are readily available from the vision community to both detect and track the required affordance-centric frames. These models typically exhibit open-world generalisation with the ability to operate over a vast range of object instances. A complete visual overview of our perception pipeline is given in Figure 1. For a given task, we first identify the object that the robot will be operating on and pass this text description together with an image of the scene as input to Grounding DINO [49]. This model provides a bounding box around the object of interest. We pass this bounding box to SAM [25] to obtain a segmentation mask of the object in the image which we use to initialise Foundation Pose [24]. For Foundation Pose, we additionally require a textured mesh of the object of interest. While there are several ways in which this can be obtained we found CSM Cube's [50] Image-to-3D model worked well in most cases where the object exhibited a repeat pattern or uniform colour in all directions. We appropriately scale this mesh within CSM Cube. Using the mask and textured mesh as input, we obtain a pose estimate of the object using Foundation Pose which is centred at the origin of the input mesh. To transform this frame to the affordance region of the object, we utilise DINO-ViT [51] to retrieve the centroid of the affordance region using a stored set of reference points for that particular object. The affordance frame is obtained by simply translating the object pose frame to this region. Once initialised, we continue to track this frame using Foundation Pose at 20Hz.

### 4.2 Oriented-Affordance Frame

Learning a policy relative to a reference frame is conceptually straightforward but challenges arise when the frame is subject to free rotation during interaction. In our context, this issue becomes particularly pronounced as we deal with continuous closed-loop interaction with objects, as opposed to one-step, open-loop tasks like grasping. The reference frame can change dynamically throughout the task, especially in non-prehensile manipulations such as pushing. These rotational changes are problematic when using a fixed-base robot, as unrestricted rotations can lead to undesirable pose configurations, potentially causing violations of joint limits. To address this, we introduce the concept of an oriented affordance frame, which stabilises the relative frame during manipulation by reorienting its x-axis (funnel) to always point towards the starting pose of the robot's tool frame. While the affordance frame can still translate with the object as it is manipulated, its orientation is anchored towards a fixed spatial point. This stabilisation reduces joint limit violations and simplifies data collection to collecting demonstrations with the object fixed in place and rotated as shown in Figure 3. The broader implications of this anchoring strategy, particularly its contribution to the robustness and success of manipulation tasks, are discussed further in Section 6.

**Algorithm 1:** Calculation of ${}^{\text{o-aff}}\mathbf{R}_{\text{afford}}$

**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** $\mathbf{R}_{align}$

**State and Action Representation:** We define a given task by its corresponding affordance and tool frame transforms as follows: $\mathcal{T}_i = \{\mathbf{T}_{\text{afford}}, \mathbf{T}_{\text{tool}}\}$ where $\mathbf{T}_{\text{afford}} \in \mathbf{SE}(3)$ and $\mathbf{T}_{\text{tool}} \in \mathbf{SE}(3)$. We use these transforms to obtain our oriented affordance frame $\mathbf{T}_{\text{o-aff}} \in \mathbf{SE}(3)$ using Algorithm 1.

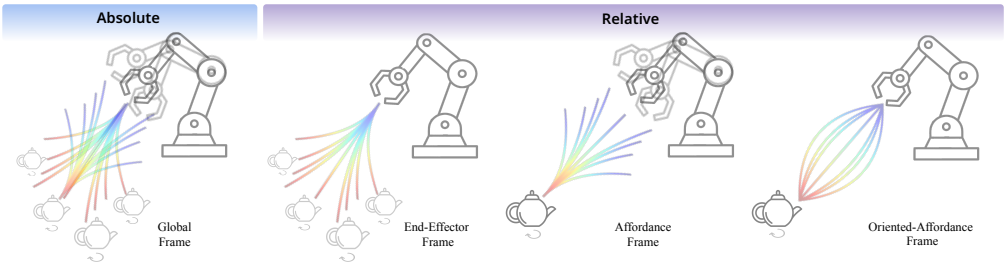



Figure 3: **Demonstration trajectory distributions for different frames of reference.** *Absolute:* A *fixed global* reference frame requires all spatial arrangements of both the end-effector and the object. *Relative:* An end-effector or affordance-centric reference frame only requires one of them to freely translate relative to the other. An *oriented*-affordance frame of reference only requires the object to freely rotate in one location to capture all downstream spatial variations.

We define the observation space for the corresponding policy $\pi_i$ as the pose of the tool-frame, ${}^{\text{o-aff}}\mathbf{T}_{\text{tool}} \in \mathbf{SE}(3)$, represented in the oriented affordance frame, the rotation of the affordance frame in this oriented-affordance frame ${}^{\text{o-aff}}\mathbf{R}_{\text{afford}} \in \mathbf{SO}(3)$ and the gripper state $g_{\text{s}} \in \{0, 1\}$. The action space of the policy consists of the desired next pose of the robot's end effector ${}^{\text{o-aff}}\mathbf{T}_{\text{ee}}$ in the oriented affordance frame, and the gripper action $g_{\text{a}} \in \{0, 1\}$.

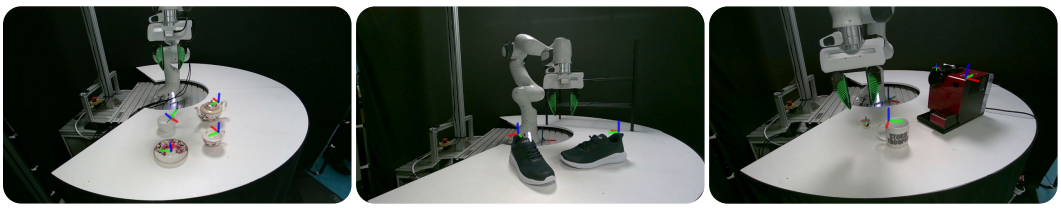

Figure 2: **Affordance Frames, Oriented-Affordance Frames and Tool Frames.** Left: Affordance frames ( ), oriented affordance frames ( ), and tool frame ( ) for a typical *pick* task. Right: Frames for the *pour* task. Notice how the oriented affordance frames are identical to the affordance frames, but rotated such that the 'funnel' axis (brown) points towards the origin of the tool frame at the beginning of the task.

## 5 Experiments

For any given manipulation task we can decompose the task into a series of affordance-centric sub-tasks, where the policy is trained to act within a local affordance frame. This compositionality of affordance-centric polices allows us to solve long-horizon tasks by chaining a series of affordance-centric policies. To this end, we focus our experiments and demonstrations across 3 different long-horizon, real-world tasks (Figure 4) that exhibit a series of affordance-centric sub-tasks. We describe each task below:

**1) Tea Serving:** This task involves 7 sub-tasks and 5 different objects including a teacup, saucer, teaspoon, teapot, and sugar basin. This task requires non-prehensile manipulation when rotating the cup, and delicate closed-loop movements across all subtasks to ensure the real porcelain objects used would not break.

**2) Shoe Racking:** This task involves 6 sub-tasks and 3 different objects including a left shoe, a right shoe, and a shoe rack. This task requires non-prehensile manipulation to push the shoes together and precise multi-object grasping to pick the two shoes up together.



Tea Serving | Shoe Racking | Coffee Making

Figure 4: **Tasks.** We demonstrate our system across 3 diverse tasks which exhibit different levels of complexity and precision to emphasise the robustness of our system.

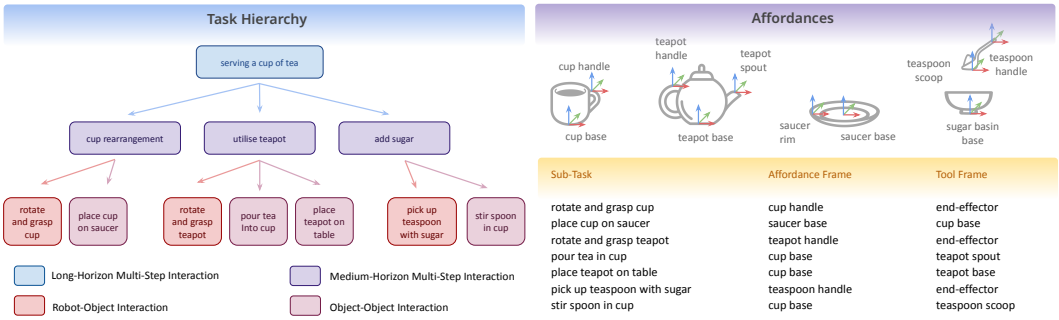



Figure 5: **Affordance-centric task decomposition for the tea serving task.** *Left:* Task decomposition hierarchy; *Top Right:* Affordance-centric frames for each object; *Bottom Right:* Sub-task frame definitions.

**3) Coffee Making:** This task involves 7 sub-tasks and 4 different objects including a coffee pod, coffee mug, coffee machine and lid. This task requires articulated closed-loop object manipulation to open/close the lid and precise placement of the pod in the machine.

### 5.1 Experimental Setup

**Task Description:** We focus our key set of evaluations on the tea-serving task. This task consists of 5 different objects including a teacup, saucer, teapot, sugar basin and teaspoon. For each object, we identify the set of affordance-centric regions which could be used across various different tasks as shown in Figure 5. Based on these affordance frames, we decompose the full tea-making task into 7 different sub-tasks. This task decomposition is done manually based on a series of unique interactions between the robot's tooltip and an affordance frame. Each unique tool-frame – affordance-frame interaction is denoted as a single affordance-centric sub-task. The full long-horizon task decomposition is given in Figure 5 (left) and we identify both the affordance and tool frames which define each subtask in Figure 5 (right). To thoroughly evaluate our ability to learn sample efficient and generalisable sub-policies and our ability to compose these sub-policies across varying levels of long-horizon task complexity, we conduct all evaluations across the task hierarchy shown in Figure 5.

**Perception System:** To evaluate the different components of our proposed system we utilise 2 different perception systems:
*1) Marker-based:* To evaluate the utility of oriented affordance-centric task frames for behaviour cloning, we decouple our results from the performance of the perception system by utilising ground truth detection and tracking of these frames in the form of fiducial markers (ArUco) placed at the affordance centric regions on objects.
*2) Large Vision Models:* Our proposed pipeline for detecting and tracking affordance frames using a series of pre-trained large vision models as illustrated in Figure 1.

**Policy Training:** We utilise Diffusion Policy [1] for imitation learning and train each policy for 4500 epochs with the same default parameters provided in the original implementation [1]. The state space for all the affordance-centric policies comprises a 16-D vector consisting of the 3-D position of the robot's tool frame, a 6-D representation [52] of the tool frame and object rotation, and the 1-D gripper state. For all the baselines, we additionally provide the position of the object resulting in a 19-D state vector. The action space for all methods is the same and comprises a 11-D vector consisting of the 3-D position of the robot's end effector, a 6-D representation [52] of its rotation, the 1-D gripper action and the 1-D policy's self-progress.

**Evaluation Methodology:** A key goal of this paper is to explore how affordance-centric policy learning enables both better sample efficiency and generalisation for manipulation tasks. To this end, we limit our training of all sub-policies to only 10 demonstrations. This constraint allows us to better understand the generalisation capabilities of our method in the low data regime while trying to decouple its success from simple data scaling. We consider both in-domain (IND) and out-of-distribution (OOD) task evaluations. In (IND) evaluation, the policy is evaluated in regions and object arrangements it was explicitly trained on, whereas in (OOD) evaluation, we evaluate its spatial and compositional generalisation to new object and inter-object locations. For each task,

| Task | # of Demos | Oriented Affordance Frame (Ours) | | End Effector Frame | | Global Frame | |
|---|---|---|---|---|---|---|---|
| | | IND Success | OOD Success | IND Success | OOD Success | IND Success | OOD Success |
| rotate and grasp cup | 10 | 81.8% | 81.8% | 45.5% | 45.5% | 45.5% | 0.0% |
| place cup on saucer | 10 | 100% | 100% | 100% | 100% | 9.1% | 0.0% |
| rotate and grasp teapot | 10 | 90.9% | 81.8% | 27.3% | 27.3% | 81.8% | 0.0% |
| pour tea into cup | 10 | 100% | 81.8% | 45.5% | 27.3% | 54.5% | 0.0% |
| place teapot on table | 10 | 90.9% | 72.7% | 54.5% | 54.5% | 90.9% | 0.0% |
| pick up teaspoon with sugar | 10 | 81.8% | 81.8% | 45.5% | 27.3% | 72.7% | 0.0% |
| stir spoon in cup | 10 | 90.9% | 81.8% | 18.2% | 9.1% | 72.7% | 0.0% |

Table 1: **Sub-policy evaluation.** Success rate across both in-distribution (IND) and out-of-distribution (OOD) spatial configurations of objects for each sub-task.

| Task | # of Demos | Oriented Affordance Frame (Ours) (Composition) | | | Global Frame (Composition) | | | End Effector Frame (Composition) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | State Dim | IND Success | OOD Success | State Dim | IND Success | OOD Success | State Dim | IND Success | OOD Success |
| cup rearrangement | 10 | 16 | 81.8% | 81.8% | 19 | 36.4% | 0.0% | 19 | 45.5% | 0.0% |
| utilise teapot | 10 | 16 | 81.8% | 63.3% | 19 | 45.5% | 0.0% | 19 | 18.2 | 9.1% |
| add sugar | 10 | 16 | 72.2% | 72.2% | 19 | 63.6% | 0.0% | 19 | 9.1% | 9.1% |
| **serving a cup of tea** | 10 | 16 | 81.8% | 63.3% | 19 | 0.0% | 0.0% | 19 | 9.1% | 9.1% |

Table 2: **Policy compositionality evaluation.** Success rate across both in-distribution (IND) and out-of-distribution (OOD) spatial configurations of multiple objects for compositional tasks.

all objects start with the same set of initial states, matched manually with reference markings. We illustrate all the object configurations used for training and evaluation in Figure 7.

# 6 Results

We summarise our evaluation of each sub-policy in Table 1 and our compositional evaluation when solving extended tasks in Table 2. Across all evaluations, our oriented affordance frame consistently outperforms all alternative methods across both learning individual sub-tasks and when composing these policies to solve long horizon tasks with an average success rate of 83.1% in the (OOD) for each individual sub-policy and 70.2% in the compositional setting. We note here that all evaluations are focused on the low data regime with each policy only trained with 10 demonstrations, however, the overall performance of our method could be significantly increased to near 100% success across all sub-tasks by increasing the number of demonstrations to only >30 as shown in Figure 6.

### 6.0.1 Key Findings

**i) Sample efficiency** As shown in Table 1, across all tasks, the affordance-based policy demonstrated the highest task performance when compared to both a relative and global frame. The task oriented nature of this frame allows the demonstrations to be concentrated around critical regions for manipulation success which the other frames do not induce without significant data overhead. As shown in Figure 6 we contrast the ability of our method to learn a spatially invariant policy from the equivalent of just 10 demonstrations for the cup rearrangement task when compared to a similar image-based policy [47] which required training on 305 demonstrations[1].

| | 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% |
| Large Vision Models | **70%** | 0.0% | 66.7% | 33.3% |

Figure 8: **Abalation Study.** Analysis of failure modes when comparing the two different affordance frames and perception systems.

**ii) Spatial generalisation** Figure 7 illustrates the start configurations used for Training each policy and the OOD start state spatial variations used during Evaluation. Across all this unseen spatial variations, our affordance-based policy was able to achieve a success rate of atleast 80% which the global frame-based policy failed entirely in this setting.

[1]We attempted to follow a similar evaluation protocol as per [47] with a broader set of spatial variations.

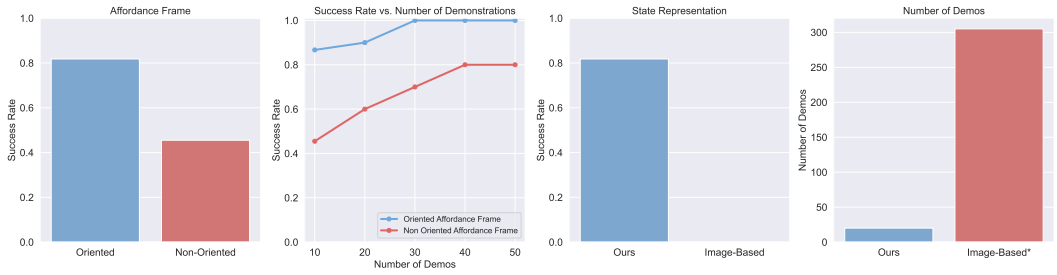



Figure 6: **Additional Comparisons.** *a)* Comparison on different affordance frames; *b)* Success rate vs. number of demonstrations for the different affordance frames; *c)* Performance of an image-based RGB policy when trained with only 10 demonstrations for the cup rearrangement task; *d)* Relative number of demonstrations required for standard image-based diffusion policy [47] to achieve the same generalisation and performance as our system.

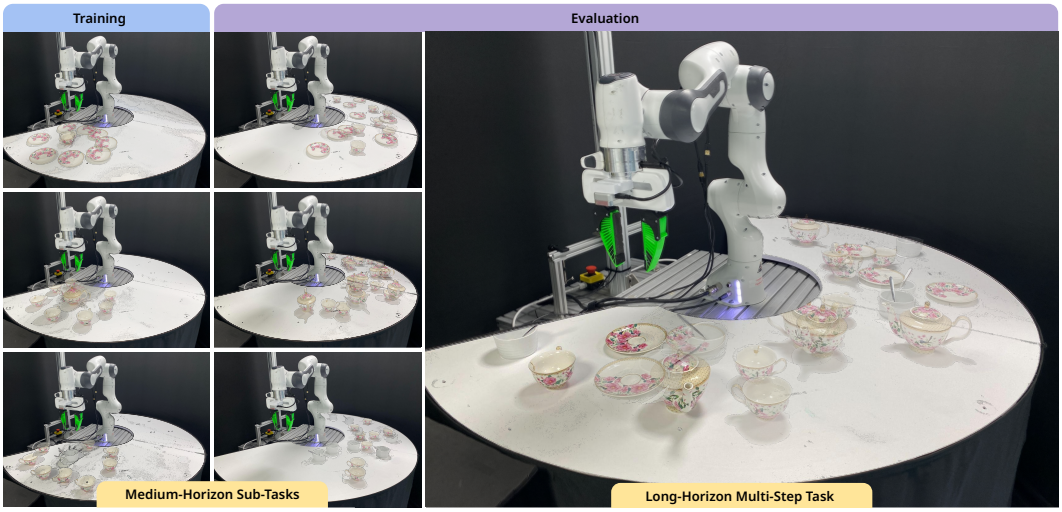



Figure 7: **Training and OOD Evaluation Object Start Configurations.** Spatial start configurations of objects across tasks used for training and evaluation in the out-of-distribution setting.

**iii) Oriented Affordance Frame Simplifies Data Collection** By orienting the affordance frame towards the start location of the robots tool-frame we can structure data collection to a small region of the work space as shown in the left column of Figure 7 while still being able to operate across the entire robot task space at test time. Furthermore, the need for only 10 demos, and the low dimensionality of the state allowed us to collect demonstrations for a task in 15 minutes and have a diffusion policy trained within the next 20 minutes enabling faster training-evaluation cycles for behaviour cloning.

**iv) Reduced joint limit violations** While relative action frames provide the ability to generalise policies to new spatial configurations, we found that the end effector might rotate to achieve a desired pose relative to the object without considering the robot's base orientation. This often led to awkward and constrained configurations, resulting in joint limit violations. These violations were a common occurrence that led to failed trials when evaluating the end-effector or affordance-centric baseline, particularly in tasks where the robot needed to rotate the object


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

Figure 9: **Generalisation to intra-category variations** The set of objects used for training and evaluating the intra-category generalisation capabilities of the trained sub-policies.

as shown in Figure 6 (left) and Table 8. Our key insight from these observations is the critical importance of the *oriented* affordance frame. By anchoring the movements of the end-effector relative to the object's affordance frame, the robot can avoid excessive rotations and unwanted configurations.

**v) Intra-category invariance for imitation learning** By attaching our relative frame for imitation learning at the affordance-centric regions of an object, we gain the ability to transfer our trained policy across a wide range of intra-category variations which share the same affordances. We illustrate this in Figure 9 where we train both the cup rearrangement and teapot utilisation tasks on a single cup and teapot set as shown in the right panels. The same trained policy was evaluated across a wide range of variations ranging from colour, shape and size, with each policy achieving almost a perfect success rate. Allowing the tool-frame state of the policy to vary based on the object's shape played an important role in generalising the policy to larger intra-category variations where the cup was significantly smaller or the spout of the tea-pot retracted significantly as shown in the bottom right of Figure 9.

**vi) 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 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.

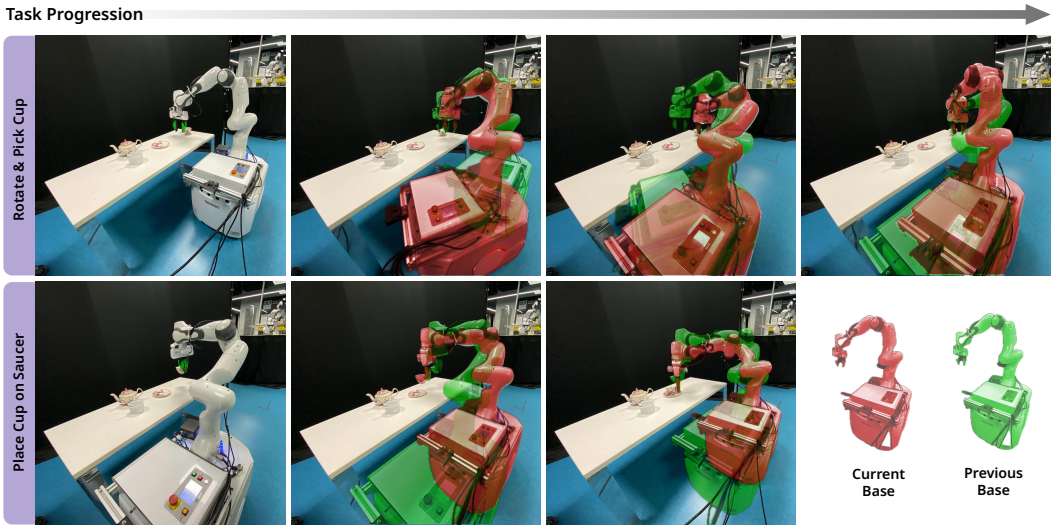



Figure 10: **Robustness to moving base.** We demonstrate our ability to maintain task performance regardless of the robot's moving base when operating with respect to an affordance-centric task frame.

**vii) Offloading Generalisation to Large Vision Models** Overall, we demonstrate that the generalist capabilities of large vision models enable us to obtain effective state abstractions for policy learning without the need for custom, narrow perception modules. By appropriately sequencing these models (Figure 1) we were able to obtain performance close to that of ground-truth perception systems as shown in Table 8. This demonstrates that the ability to circumvent the need for extensive robot data can be achieved by appropriately leveraging the existing generalist vision systems already available.

## 7 Conclusion

We propose an affordance-centric policy learning framework and a general pipeline that detects and tracks affordances using pre-trained large vision models. Affordances offers significant advantages, including invariance to robot and object poses, robustness to task-irrelevant visual attributes, and flexibility in behaviour composition across various scene configurations. To achieve these invariances while ensuring reliable deployment across diverse object configurations, we introduce the oriented affordance frame as an effective method for anchoring relative task frames. Experimental

results show substantial improvements in sample efficiency and generalisation, significantly simplifying the data requirements for generalisable behaviour cloning.

**Limitations:** Our system still faces several limitations. Our reliance on perception systems means that policy performance depends heavily on their robustness in real-world scenarios, particularly for tracking under clutter, where current methods still face challenges. Future work could explore advanced 3D tracking techniques [53] to mitigate these issues. Furthermore, our approach is not directly applicable to non-rigid objects, requiring additional state information. The pose-based abstraction may also limit its application to tasks requiring finer details, necessitating additional sensory modalities such as tactile sensing. Despite these limitations, our method represents a significant step toward more efficient and generalisable behaviour cloning for complex manipulation tasks.

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

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[4] 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.

[5] 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.

[6] J. Park, Y. Seo, C. Liu, L. Zhao, T. Qin, J. Shin, and T.-Y. Liu. Object-aware regularization for addressing causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 34:3029–3042, 2021.

[7] F. Codevilla, E. Santana, A. M. López, and A. Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9329–9338, 2019.

[8] P. De Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. *Advances in neural information processing systems*, 32, 2019.

[9] I. Kostrikov, O. Nachum, and J. Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. *arXiv preprint arXiv:2312.08344*, 2023.

[25] 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.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[27] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[28] T. B. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[29] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

[31] 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.

[32] H. Huang, D. Wang, R. Walters, and R. Platt. Equivariant transporter network. *arXiv preprint arXiv:2202.09400*, 2022.

[33] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt. Sample efficient grasp learning using equivariant models. *arXiv preprint arXiv:2202.09468*, 2022.

[34] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR, 2021.

[35] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[36] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

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

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

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

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

[51] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

[52] 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.

[53] J. Abou-Chakra, K. Rana, F. Dayoub, and N. Sünderhauf. Physically embodied gaussian splatting: Embedding physical priors into a visual 3d world model for robotics. In *Workshop on Neural Representations for Robotics at Conference on Robot Learning*, number 7th, 2023.