AffordGen: Generating Diverse Demonstrations for Generalizable Object Manipulation with Affordance Correspondence

Jiawei Zhang[1]*, Kaizhe Hu[2,1]*, Yingqian Huang[1,3], Yuanchen Ju[4], Zhengrong Xue[2,1], Huazhe Xu[2,1]†

[1]Shanghai Qi Zhi Institute [2]Tsinghua University [3]Fudan University [4]UC Berkeley

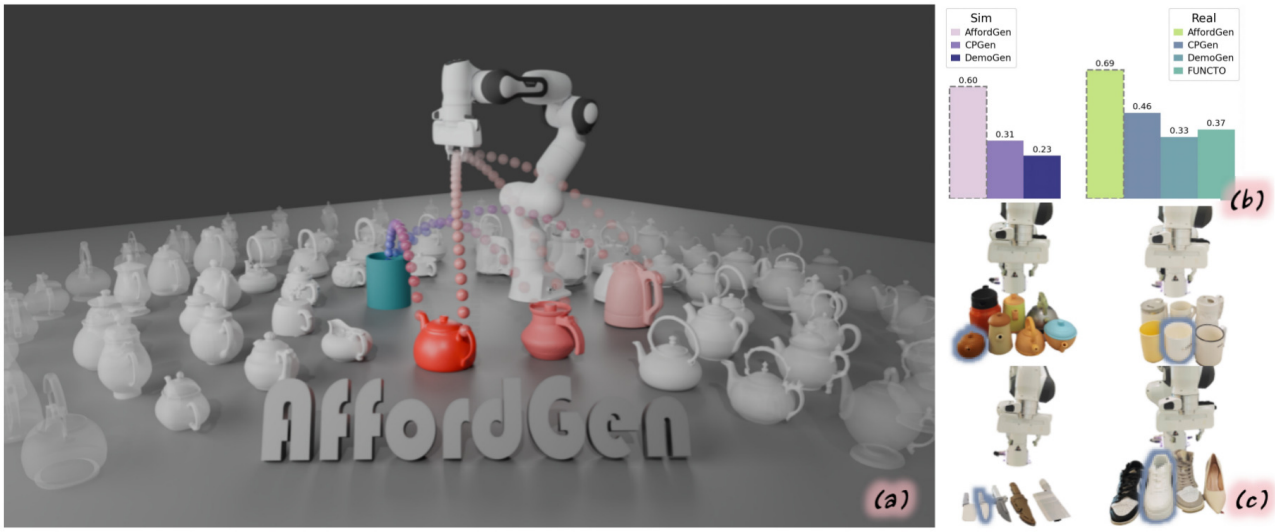


Figure 1. **AffordGen overview.** (a) Diverse trajectory generation for novel objects via one-shot demonstration. (b) Superior performance against powerful baselines. (c) Real-world generalization to unseen objects from a single source.

Abstract

*Despite the recent success of modern imitation learning methods in robot manipulation, their performance is often constrained by geometric variations due to limited data diversity. Leveraging powerful 3D generative models and vision foundation models (VFMs), the proposed **AffordGen** framework overcomes this limitation by utilizing the semantic correspondence of meaningful keypoints across large-scale 3D meshes to generate new robot manipulation trajectories. This large-scale, affordance-aware dataset is then used to train a robust, closed-loop visuomotor policy, combining the semantic generalizability of affordances with the reactive robustness of end-to-end learning. Experiments in simulation and the real world show that policies trained with AffordGen achieve high success rates and enable zero-shot generalization to truly unseen objects, significantly improving data efficiency in robot learning.*

1. Introduction

Visuomotor imitation learning has achieved impressive progress in robotic manipulation [2, 10, 12, 28, 29, 36]. However, its practical deployment is hindered by two fundamental challenges: the prohibitive cost of collecting large-scale, high-quality human demonstrations, and the poor generalization of learned policies to novel objects and scenarios not encountered during training [35]. This reliance on extensive demonstration constrains the applicability of imitation learning in diverse real-world applications.

Synthetic data generation offers one promising remedy. Recent works like DemoGen [33] can expand a single demonstration into hundreds of spatially diverse trajectories, greatly improving data efficiency. Nevertheless, these approaches face a critical limitation: they primarily augment spatial relationships for a *single* object instance. As a result, they inherit the restricted semantic scope of the source demonstration and generalize poorly beyond the

source object. They also focus more on the translation invariance of the given task, exhibiting a limited ability to deal with different orientations.

Affordance-based methods, such as Robo-ABC [8] and DenseMatcher [39], leverage semantic correspondence to transfer affordance knowledge to unseen objects. They can transfer the manipulation method of one instance of an object to the other, enabling one-shot imitation learning methods with cross-category generalization, such as FUNCTO [23] possible. However, these approaches are typically planning-centric, and rely heavily on the accuracy of the mapped affordance point and the planning algorithm. The execution of the policy just follows the precomputed, open-loop trajectories, lacking the reactive ability of learning-based closed-loop policies. Thus, while affordance research is highly active, a systematic approach to effectively integrate this semantic knowledge into learning-based pipelines is still missing.

In this work, we introduce AffordGen, a novel framework that repurposes affordance information as a generative prior for policy learning. AffordGen produces feasible training data on unseen objects that may come from a different category. Starting from a few human demonstrations, we establish correspondences between the demonstrated object's keypoints and a large set of unseen 3D models. Then, we synthesize diverse yet semantically grounded trajectories in the style of DemoGen, but crucially over novel object instances and full 6D spatial relations. In this way, AffordGen scales a handful of demonstrations into thousands of high-quality, affordance-aware trajectories that cover a diverse realm of objects.

The key innovation of AffordGen lies in its novel use of affordance knowledge to generate trajectories. We argue that affordance knowledge is best utilized not through online planning, but as a powerful guide of data generation. By leveraging affordances to generate a large, diverse dataset of plausible trajectories, we can then train a reactive, closed-loop visuomotor policy that inherits both the semantic generalizability of affordances and the robustness of end-to-end learning.

AffordGen demonstrates impressive performance on both simulation and real-world tasks, achieving an average performance boost of **24.1%** and **24.3%** over the best baseline on simulation and real tasks. It can generate thousands of meaningful trajectories over hundreds of different meshes, achieving a high success rate from as few as one source demonstration. AffordGen boosts the model performance on both generated "seen" objects and truly unseen objects that are not in the generated dataset. Beyond these abilities, it also has the potential to generate new data on completely different objects, as long as they share the same manipulation type. We also find that the object-level generation ability increases first and then decreases as we extend the generation to more unseen objects, guiding future cross-instance generation works. The contribution of AffordGen can be summarized below:

- We introduce AffordGen, a novel framework that repurposes affordance correspondence as a generative source, enabling the synthesis of diverse and semantically meaningful demonstrations.
- We demonstrate that our method can scale a minimal number of human demonstrations into thousands of trajectories across novel object categories and full 6D poses, overcoming the semantic and geometric limitations of prior data augmentation techniques.
- We validate through extensive experiments that policies trained with AffordGen achieve significant zero-shot generalization to unseen objects, presenting a new paradigm for data-efficient robot learning.

2. Related Works

2.1. Data Generation for Robot Manipulation

Automated data generation techniques for robot manipulation have great potential to alleviate the data problem of embodied AI. These methods primarily fall into two categories: adapting existing demonstrations and generating new data from scratch.

The first paradigm, trajectory adaptation, is exemplified by MimicGen [14] and its variants [4, 7, 16], which segment a few source demonstrations into object-centric skills and replay them in new spatial configurations. While powerful, these methods rely on costly on-robot or in-simulation rollouts to capture the final interaction data. DemoGen [33] addresses this bottleneck by introducing a fully synthetic pipeline that generates paired actions and observations from direct 3D point cloud editing. More related to our approach, CPGen [11] extends DemoGen by stretching and transforming the source mesh to improve the diversity of generated data and the generalization ability of the policy. These methods, however, are still limited to the source object and have poor generalization ability to unseen objects even within the same category.

The second paradigm leverages large-scale generative models to create data from scratch. Systems like GenSim [30], GenSim2 [6], and RoboGen [31] use Large Language Models (LLMs) to propose novel tasks, generate corresponding scenes, and script automated solvers to create demonstrations. This approach excels at generating task-level diversity but is limited by the capability of the underlying automated solvers, which may not match the quality of human demonstrations. Orthogonal to this, another line of work [1, 5, 13, 34] uses diffusion models to augment the visual appearance of existing data, enhancing a policy's visual robustness but not its ability to generalize to new spatial configurations or object types.

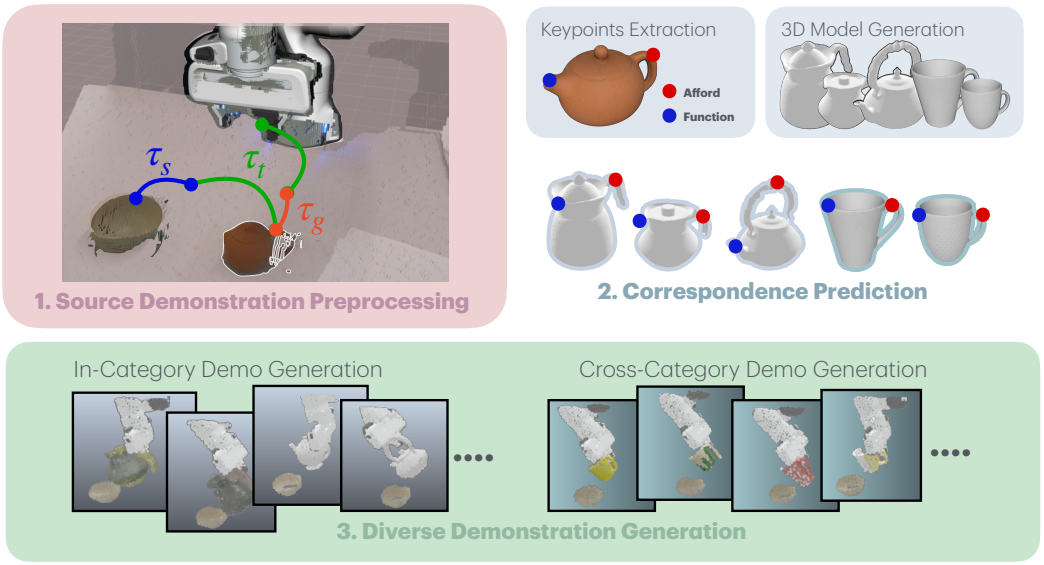


Figure 2. 1. AffordGen takes in a source expert demonstration and splits it into different functioning segments. 2. We extract keypoints on the source object and establish correspondences between them and many target objects in 3D space through visual foundation models. 3. By transferring the task-related segments and planning the transition segments, AffordGen can generate diverse and large-scale trajectories on both in-category and cross-category objects.

2.2. Semantic Correspondence for Manipulation

Early research on semantic correspondence for manipulation relies on visual descriptors and semantic keypoints, such as DON [3], kPAM [15], and NDF [21], which enabled instance-level transfer by aligning consistent geometric or appearance features. Current works exploit foundation models like diffusion models [25, 37] for semantic alignment, and recent studies have extended such capabilities to robotic manipulation by leveraging semantic correspondence for affordance transfer across objects and categories [8, 9, 23, 32, 39]. Beyond these, other lines of research further explore semantic correspondence in different contexts: MimicFunc [24] transfers functional correspondences from human videos to novel tools, and HRP leverages human affordances for robotic pre-training [22].

Although these methods expand the scope of semantic correspondence, most still treat affordances merely as mapping signals, using them to locate contact points and relying on planners for execution, thus lacking true reactivity and flexibility. In contrast, AffordGen transforms affordances from static mappings into generative sources, synthesizing large-scale, diverse, and semantically consistent demonstrations that not only span across instances and categories but also provide rich data for training visuomotor policies, thereby achieving stronger generalization and adaptability.

3. Methodology

3.1. Problem Formulation

Like many prior online planning works [18, 23] for robotic manipulation, we decompose a manipulation task into three distinct stages $\Omega = \{\Omega_G, \Omega_S, \Omega_T\}$, where Ω_G denotes the grasp stage during which the robot closes its gripper to secure the object. Ω_S denotes the skill stage where the robot manipulates the grasped object to accomplish the task, like pouring tea into the tea cup, and Ω_T denotes the transition stage that connects the two other stages without collision. At each timestep t within a stage, the robot takes in its current visual observation o_t^e and proprioception observation o_t^s, and outputs a corresponding action a_t. AffordGen uses point clouds as the input type of o_t^e because of its simplicity and special structure in 3D space that can be easily manipulated to generate new data.

The generation strategy of AffordGen is inspired by the intuition shared across the grasp and skill stages: despite the shapes and sizes of the objects manipulated in a given task varying, the semantic information embedded in their trajectories remains similar. In the following sections, we will discuss: (1) how to extract semantic information from the grasp and skill stages in the original demonstrations; (2) how to map this information onto large-scale 3D meshes different from the source; and (3) how to generate a large

set of diverse demonstrations from the source trajectory.

3.2. Source Demonstration Pre-processing

Given an expert demonstration, we consider extracting three types of information: the *grasping time* t_{grasp} at which the gripper closes during Ω_G; the *skill segment* τ_s where the actions crucial to the task's success are performed; and the keypoints of the manipulated object in the original data, namely the *affording point* and the *function point*.

The grasping time t_{grasp} can be directly extracted according to the end-effector state information from the expert demonstration. The skill segment τ_s can be detected either through the video reasoning capabilities of vision-language models (VLMs) or by direct human annotation. The *affording point* refers to the contact point between the gripper and the object, while the *function point* refers to the point the tool interacts with other objects to accomplish the task. Both points are defined in 3D space, and can be readily recognized and annotated by VLMs or humans. As we only have a limited number of demonstrations, this process remains straightforward and efficient.

The point cloud used as visual input is derived from an RGB-D camera. On the raw RGB images, we employ SAM2 [19] to segment the objects involved in the task into four categories: robot, object, goal, and others. The segmentation labels on the 2D images are then mapped onto the point cloud along with the pixel coordinates, resulting in a segmented point cloud. Following the settings of DP3 [36] and DemoGen, we further remove the background and floor from the obtained point cloud and then use Farthest Point Sampling (FPS) to down-sample it, thereby extracting the workspace point cloud $\mathcal{O}^e \subset \mathbb{R}^3$, which contains only the points located within the robot's workspace.

3.3. Semantic Correspondence on 3D Meshes

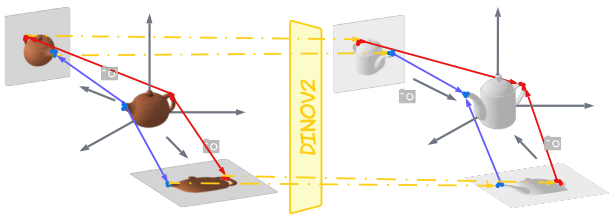


Figure 3. Keypoints Correspondence in 3D Canonical Space. The keypoints are mapped to the target mesh through DINOv2.

To enable the reuse of trajectory information from the source demonstration, we need to establish correspondences between the keypoints of the manipulated object in the source demo and those on the new 3D meshes. While semantic correspondence across 2D images has been extensively studied, robotic manipulation tasks require accurate annotations in 3D space. Recent works have begun to explore semantic correspondences across 3D meshes [39], but these approaches have been trained only on small-scale datasets, limiting their applicability to precise robotic manipulation tasks. To address this limitation, we propose a simple yet effective approach that normalizes all 3D meshes into a unified canonical space before mapping the 2D correspondences onto the 3D meshes.

Formally, let the source mesh be denoted as M_{src}, and let its associated keypoint be $x \in \mathbb{R}^3$, which lies in the local frame of M_{src}. Our objective is to find the corresponding keypoint x' on the target mesh M_{tg}. For any given mesh, its pose is first normalized into the canonical space using a 6D pose estimator [38]. We then perform parallel rendering to obtain RGB-D images from n different camera views. The pose of the i-th camera is denoted as P_i $(i = 1, 2, \ldots, n)$, and the corresponding rendered image as I_i. Each image I_i is fed into DINOv2 [17] to obtain its semantic representation S_i. We select m nearest mesh vertices in the neighborhood of x, denoted as v_j $(j = 1, \ldots, m)$. Each vertex v_j is projected onto the image I_i via forward camera projection, resulting in the pixel coordinate u_{ij}. The corresponding pixel of v_j on the target image is obtained by maximizing the cosine similarity in the feature space of the DINOv2 model:

$$u_{ij}^{\text{tg}} = \arg\max_{u} \ \text{CosSim}\left(S_i^{src}[u_{ij}],\ S_i^{tg}[u]\right),$$
$$w_{ij} = \text{CosSim}\left(S_i^{src}[u_{ij}],\ S_i^{tg}[u_{ij}^{\text{tg}}]\right).$$

All the matched pixels u_{ij}^{tg} are then unprojected into the 3D space, yielding candidate correspondences v_{ij}^{tg} with associated similarity scores w_{ij}. Finally, the target keypoint x' in 3D space is obtained by weighted average:

$$x' = \frac{\sum_{i,j} w_{ij}\, v_{ij}^{\text{tg}}}{\sum_{i,j} w_{ij}}.$$

3.4. Diverse Demonstration Generation

We follow a three-step approach to generate demonstrations on new meshes. We consider a robotic manipulation scenario where the end-effector manipulates an *active object* to interact with a *goal object*. In this framework, the *active object* is defined as the object being directly maneuvered by the robot (e.g., a teapot), while the *goal object* is the object that the active object acts upon (e.g., a cup).

3.4.1. Keypoint-constrained Trajectory Replay

Our goal is to efficiently transfer the grasp segment τ_g and the skill segment τ_s from source demonstrations to novel objects. We leverage the correspondence between the affording points $(x_{\text{aff}}, x'_{\text{aff}})$ and the function points $(x_{\text{fun}}, x'_{\text{fun}})$, following a simple yet effective assumption: when manipulating objects of the same *function class*, the trajectories of

the end effector relative to the affording point remain similar, while the trajectories of the function point relative to the goal object remain similar as well. The *function class* here extends the definition of object categories and refers to objects that share similar functionality. For example, a mug and a teapot share the same function of pouring water into a cup. An illustration is shown in Figure 4.

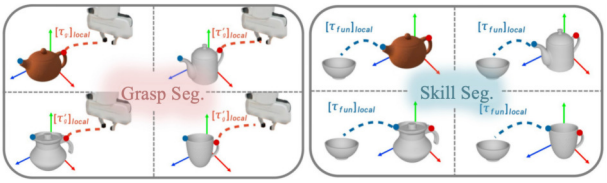


Figure 4. Trajectory replay for grasp and skill segments. τ_g and τ_s remain similar across all meshes.

To transfer **the grasp segment** τ_g, we first normalize the end-effector segment τ_g into the local frame of the source mesh by using the initial pose of the source object T_{init}. The normalized grasp segment can be expressed as $[\tau_g]_{\text{local}} = T_{\text{init}}^{-1} \cdot [\tau_g]_{\text{world}}$. Given a new object mesh M_{tg} with the associated affording point x'_{aff}, the new grasp segment in the local frame of M_{tg} can be obtained by

$$[\tau'_g]_{\text{local}} = [\tau_g]_{\text{local}} - x_{\text{aff}} + x'_{\text{aff}}$$

To transfer **the skill segment** τ_s, we first derive the trajectory of the function point during the skill stage Ω_S in the world frame by leveraging its relative transformation to the end effector $T_{\text{ee}}^{\text{fun}}$. Thus, the function point trajectory in both world and local frame can be expressed as

$$[\tau_{\text{fun}}]_{\text{world}} = T_{\text{ee}}^{\text{fun}} \cdot [\tau_s]$$
$$[\tau_{\text{fun}}]_{\text{local}} = T_{\text{init}}^{-1} \cdot [\tau_{\text{fun}}]_{\text{world}}$$

respectively. For the new mesh M_{tg} and its function point x'_{fun}, the new function trajectory in the local frame of M_{tg} can be obtained by

$$[\tau'_{\text{fun}}]_{\text{local}} = [\tau_{\text{fun}}]_{\text{local}} - x_{\text{fun}} + x'_{\text{fun}}$$

For any random pose configuration T' of mesh M_{tg}, the new grasp segment and skill segment can be solved from $[\tau'_g]_{\text{local}}$ and $[\tau'_{\text{fun}}]_{\text{local}}$ by

$$[\tau'_g] = T' \cdot [\tau'_g]_{\text{local}}$$
$$[\tau'_s] = T_{\text{fun}}^{\text{ee}} \cdot T' \cdot [\tau'_{\text{fun}}]_{\text{local}}$$

The resulting τ'_g and τ'_s correspond to the sequence of end-effector poses in world frame for executing the skill on the new mesh. The joint positions for each waypoint can be further computed via an inverse kinematics (IK) solver.

3.4.2. Motion Planning for Transition Segment

In most robot manipulation tasks, the connection segment between the grasp segment and the skill segment contains much less meaningful information. Such trajectories do not involve dynamic interactions between the robot arm and the active object, nor between the active object and the goal object. Therefore, it can be regarded as a collision-free free-space motion between the grasping segment and the skill segment. We employ motion planning to plan this trajectory or utilize spherical linear interpolation [20] to directly interpolate the path. We denote the resulting trajectory as τ_{m_i}, where i represents the i-th free-motion segment during the task. For example, the free motion segment between τ'_g and τ'_s can be expressed as

$$\tau'_m = \text{MotionPlan}\left(\tau'_g[-1], \tau'_s[0]\right)$$

3.4.3. Point Cloud Digital Cousin Generation

After obtaining trajectories τ for novel objects, it is necessary to transform the point cloud of the source data accordingly to align with these new trajectories and meshes. DemoGen[33] achieves the generation of point clouds for new trajectories by applying simple global translation and rotation on the original point cloud. This approach is effective because the trajectories generated by DemoGen differ from the original ones primarily by a positional shift. In contrast, AffordGen aims to produce a variety of 3D models with diverse 6D poses within the workspace. To achieve this, we directly render the robot and the manipulated object point clouds from simulation, and then replace their counterparts in the source demonstration. The resulting hybrid real-simulated point cloud data mitigates the sim-to-real gap while bypassing the tedious process of fully reconstructing the environment in simulation. We employ parallel rendering to generate point clouds to achieve high throughput efficiency. A visualization of the generated trajectory, in comparison to the source demonstration, is provided in Figure 5.

Apart from these approaches, we apply special modifications of the skill segment to deal with the occlusion problem. For a detailed discussion of the data generation scheme and the generation quality, please refer to Appendix 9

4. Experiments

To validate our claims about the effectiveness of AffordGen in generating data through the transfer of grasp segments and skill segments, we conduct generalization experiments in both simulation and the real world.

4.1. Simulation Experiments

4.1.1. Experiment Settings

We use ManiSkill3 [26] as the simulator and construct four tasks involving both grasp and skill stages: *teapot pouring*,

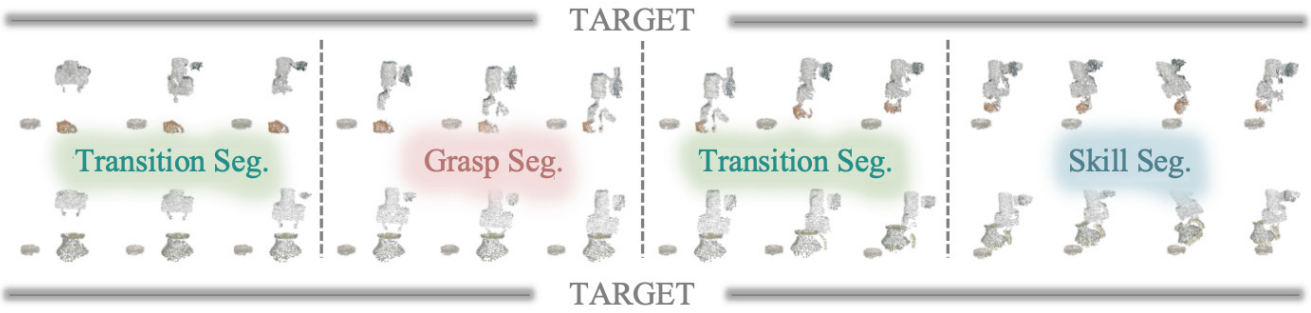


Figure 5. Visualization of source and generated trajectory of the teapot pouring task. The upper line is the source trajectory, while the lower line is a generated trajectory. Each trajectory is composed of three types of segments.

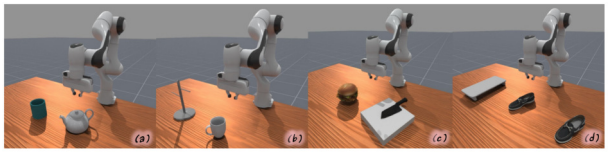


Figure 6. Simulative experiments setup: (a) Teapot Pouring, (b) Mug Hanging, (c) Knife Cutting, (d) Shoe Organizing.

mug hanging, *knife cutting* and *shoe organizing*, as illustrated in Figure 6. For *teapot pouring*, the robot needs to grasp the handle part of the teapot and then move and tilt the teapot to position its spout above the target cup. For *mug hanging*, the robot needs to grasp the mug by its handle, move it, and insert the rack prong into the hole of the mug handle. For *knife cutting*, the robot first grasps the handle of the knife and then maneuvers it to make contact between the blade and the target food. For *shoe organizing*, which is a long-horizon and multi-object task, the robot grasps the shoes at the heels and places both the nearer and the farther shoes onto the shoe rack in correct orientation.

For each task, we collect **one** expert demonstration as a reference for generation and generate **1000** demonstrations using 3D mesh assets, from the PAM [38] dataset or 3D generative model [27]. The 3D meshes are randomly split into *seen* and *unseen* subsets; the former is used for data generation and the latter for evaluation. For detailed information on the 3D mesh dataset split and visualization, please refer to Appendix 8.

We compare the performance with DemoGen [33] and CPGen [11]. Similar to AffordGen, DemoGen also operates on 3D point clouds. We use the same source demonstration and generate 1000 trajectories using DemoGen. We take DemoGen as one of the baselines to demonstrate that AffordGen maintains its spatial generalization capability on the original mesh when generalizing to unseen shapes. CPGen generates demonstrations by stretching and transforming the original object mesh, aiming to enhance shape generalization. We also generate 1000 trajectories, but stretch them randomly in the scale range of the evaluation set. Note that the original CPGen operates on RGBD data. For a fair comparison, we adapt CPGen to work with the 3D point cloud modality. The implementation details of the baselines can be found in Appendix 7.

4.1.2. In-category Generalization Results

To thoroughly examine the efficiency of AffordGen on the generalization capability of the learned policy, we perform ablation studies on each task by fixing the total number of demonstrations while varying the number of meshes used for generation and the number of demonstrations per mesh. Under each (**#meshes**, **#demos per mesh**) combination, the learned policy is evaluated on both source mesh and 5-20 unseen meshes, where each mesh is tested 50 times with randomly placed initial positions and orientations.

The detailed performance is shown in Table 1. We separate the results into two categories: "source" and "unseen", corresponding to the performance on the source mesh and the evaluation set of meshes. DemoGen, CPGen, and AffordGen all achieve spatial generalization on the source mesh, preserving the task information of the source object over the entire workspace. Despite being trained on extensive data for the source mesh, AffordGen achieves comparable performance to DemoGen. By applying stretching and compression transformations, CPGen augments the shape of the original object, leading to better performance than DemoGen on unseen object tests. AffordGen, through semantic-corresponding keypoints, significantly expands the shape diversity of the generated data, outperforming both DemoGen and CPGen across all tasks by an average of **24.1%** on unseen object tests on the best 100×10 setting.

To further demonstrate the generalization range of our method, we select 5 evaluation meshes and arrange them based on their performance in Figure 7. The performance

#Mesh × #Demo	Teapot Pouring		Mug Hanging		Knife Cutting		Shoe Aligning	
	Source	Unseen	Source	Unseen	Source	Unseen	Source	Unseen
DemoGen (1×1000)	0.933 ± 0.009	0.131 ± 0.029	0.940 ± 0.043	0.402 ± 0.036	0.490 ± 0.037	0.224 ± 0.012	0.400 ± 0.050	0.212 ± 0.025
CPGen (1000×1)	0.713 ± 0.146	0.169 ± 0.070	0.900 ± 0.056	0.502 ± 0.027	0.747 ± 0.148	0.424 ± 0.003	0.550 ± 0.007	0.266 ± 0.024
AffordGen (20×50)	0.920 ± 0.086	0.353 ± 0.103	$\mathbf{0.967} \pm 0.025$	0.683 ± 0.004	$\mathbf{0.793} \pm 0.057$	$\mathbf{0.565} \pm 0.002$	0.550 ± 0.100	0.438 ± 0.013
AffordGen (50×20)	0.927 ± 0.025	$\mathbf{0.553} \pm 0.039$	0.777 ± 0.021	0.664 ± 0.033	0.647 ± 0.068	0.535 ± 0.006	0.588 ± 0.018	0.302 ± 0.089
AffordGen (100×10)	$\mathbf{0.960} \pm 0.043$	0.519 ± 0.072	0.753 ± 0.098	$\mathbf{0.707} \pm 0.011$	0.606 ± 0.090	0.510 ± 0.001	$\mathbf{0.825} \pm 0.035$	$\mathbf{0.588} \pm 0.018$
AffordGen (1000×1)	0.700 ± 0.123	0.242 ± 0.067	0.853 ± 0.082	0.642 ± 0.018	0.607 ± 0.207	0.542 ± 0.002	0.625 ± 0.025	0.425 ± 0.175

Table 1. Comparison of different Mesh-Demo configurations across simulative tasks.

gradually drops as the evaluation mesh becomes more dissimilar from the source, but AffordGen remains high generalization performance across all meshes.

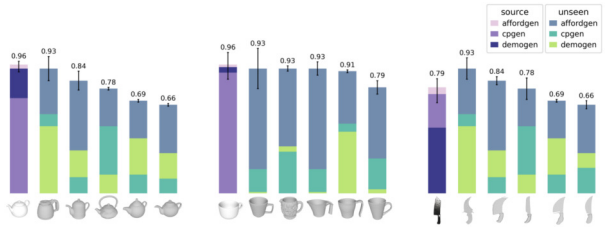


Figure 7. Simulative evaluation results on different meshes

4.1.3. Zero-shot Cross-Category Results

By establishing correspondences between keypoints, AffordGen can repurpose source demonstrations to generate training data for cross-category objects. This synthetically generated data can then be used directly to train policies on novel object categories, as long as they share the same functional affordance. To showcase such capability, we conducted 3 zero-shot cross-category policy learning experiments, namely, *mug pouring* (from teapot pouring), *handbag hanging* (from mug hanging), and *saw cutting* (from knife cutting). The simulation cross-category tasks are shown in Figure 8.

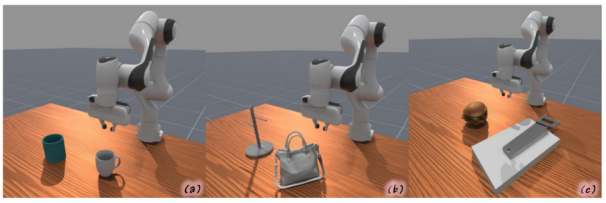


Figure 8. Simulative cross-category tasks: (a) Mug Pouring, (b) Handbag Hanging, (c) Saw Cutting.

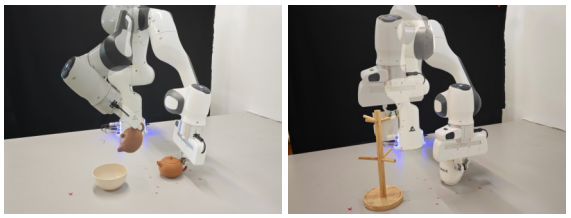
As shown in Table 3, AffordGen demonstrates impressive results in generating effective training data even for out-of-category objects. It is the only method to achieve a meaningful non-zero success rate on these new objects.

4.2. Real World Experiments

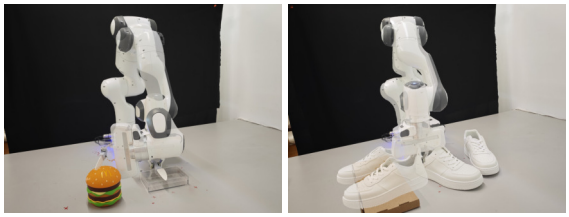
4.2.1. Experiment Settings

Real-world data generation is the most meaningful use case of AffordGen. While it is tedious and costly to collect data on thousands of different spatial relationships, it would be prohibitive to do so on thousands of different real objects.

To demonstrate the effectiveness of AffordGen in real-world generalization, we conduct the four simulation tasks in the real world. For each task, **10** expert demonstrations are collected to generate **1000** training demonstrations. Following DemoGen, we increase the number of real demonstrations to avoid overfitting. After **training solely on the generated data**, we evaluate the learned policies on a variety of previously unseen real objects.



(a) Teapot Pouring (b) Mug Hanging



(c) Knife Cutting (d) Shoe Organizing

Figure 9. Real-World experiments setup

In real-world experiments, we include another planning-based method, FUNCTO [23]. FUNCTO serves as a representative algorithm based on keypoint correspondence. Similar to AffordGen, FUNCTO generates manipulation

Table 2. Comparison of different methods across real-world tasks.

#Mesh × #Demo	**Teapot Pouring**		**Mug Hanging**		**Knife Cutting**		**Shoe Organizing**	
	Source	Unseen	Source	Unseen	Source	Unseen	Source	Unseen
DemoGen (1×1000)	**14/27**	2/162	**20/27**	74/162	**25/27**	47/108	13/20	24/60
CPGen (1000×1)	10/27	15/162	19/27	69/162	23/27	88/108	**18/20**	30/60
AffordGen (100×10)	13/27	**74/162**	19/27	**107/162**	23/27	**96/108**	11/20	**45/60**
FUNCTO	10/27	50/162	7/27	48/162	21/27	61/108	15/20	19/60

	Simulation			Real-World		
Algorithm / Tasks	Teapot-Mug	Mug-Handbag	Knife-Saw	Teapot-Mug	Mug-Handbag	Knife-Saw
AffordGen (ours)	**55.00%** ± 9.10%	**83.07%** ± 1.32%	**40.22%** ± 7.28%	**14 / 27**	**7 / 12**	**9 / 27**
CPGen	2.70% ± 2.50%	0.67% ± 0.50%	1.11% ± 1.00%	3 / 27	0 / 12	1 / 27
DemoGen	0.70% ± 0.90%	0.27% ± 0.38%	1.56% ± 0.38%	0 / 27	0 / 12	1 / 27

Table 3. Comparison of success rates on cross-category generalization tasks under simulation and real-world settings.

trajectories for new objects through semantic mapping. Leveraging large language models and visual foundation models, it performs path planning according to keypoint correspondences. We include FUNCTO to highlight the differences between AffordGen and affordance-based open-loop planning algorithms, and demonstrate how AffordGen overcomes the limitations of such approaches.

4.2.2. In-category Generalization Results

We first conduct in-category experiments on the source object and a sufficiently diverse set of unseen test objects. The task setups are illustrated in Figure 9, while the detailed evaluation setting can be found in Appendix 7.

For a fair comparison, we evaluated all test objects under a fixed set of poses: 27 poses (9 positions x 3 orientations) for most tasks, and 10 poses (2 positions x 5 orientation) specifically for the shoe task due to its long-horizon nature.

The results are presented in Table 2. We see a similar pattern to the simulation tasks: with only a small amount of real-world expert data, DemoGen, CPGen, and AffordGen all achieve high success rates across the entire workspace on the source mesh, while CPGen significantly outperforms DemoGen in most unseen tests, and AffordGen further beats both baselines.

FUNCTO's performance heavily depends on the selection of correspondence points, which is often compromised by factors such as occlusion of key parts and large view perspective differences between source and target objects. When the key regions of the target object remain clearly visible, such as knife cutting, FUNCTO maintains a relatively high success rate. However, in scenarios with large orientation variations and occlusion, such as mug hanging, FUNCTO yields the worst performance among all baseline methods. Benefiting from the 3D space correspondences and training on large-scale generated data, AffordGen implicitly learns the relationships between affording points, function points, and object shapes, effectively resolving the keypoint occlusion issues common in planner-based methods.

4.2.3. Zero-shot Cross-category Results

We establish a real-world setup identical to the simulation for cross-category testing, as shown in Figure 10. The experimental results in Table 3 demonstrate that AffordGen can effectively generate cross-category object manipulation data in the real world, thereby further expanding real-world robot capabilities with low cost.

Figure 10. Real cross-category tasks settings: (a) Mug Pouring, (b) Handbag Hanging, (c) Saw Cutting.

5. Conclusion

AffordGen addresses the data scarcity and generalization problem in robotic learning. This work presents a new paradigm that leverages affordance correspondence as a generative source to scale minimal demonstrations into thousands of diverse, semantically-grounded, and full 6D trajectories across object categories. Policies trained on this synthetic dataset achieve robust, closed-loop control and demonstrate strong generalization to truly unseen objects, indicating great potential for large-scale real-world applications.

Acknowlegement

This work is supported by Tsinghua University-Keystone Electrical (Zhejiang) Co., id Joint Research Center for Embodied Multimodal Artificial Intelligence(JCEMAI).

References

[1] Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation, 2023. 2

[2] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *ArXiv*, abs/2303.04137, 2023. 1

[3] Peter R Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv preprint arXiv:1806.08756*, 2018. 3

[4] Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment, 2024. 2

[5] Kaizhe Hu, Zihang Rui, Yao He, Yuyao Liu, and Pu Hua. Generalizable visual imitation learning with stem-like convergent observation through diffusion inversion. *arXiv preprint arXiv:2411.04919*, 1, 2024. 2

[6] Pu Hua, Minghuan Liu, Annabella Macaluso, Yunfeng Lin, Weinan Zhang, Huazhe Xu, and Lirui Wang. Gensim2: Scaling robot data generation with multi-modal and reasoning llms, 2024. 2

[7] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning, 2025. 2

[8] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, 2024. 2, 3

[9] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024. 3

[10] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *ArXiv*, abs/2410.18647, 2024. 1

[11] Kevin Lin, Varun Ragunath, Andrew McAlinden, Aaditya Prasad, Jimmy Wu, Yuke Zhu, and Jeannette Bohg. Constraint-preserving data generation for visuomotor policy learning, 2025. 2, 6

[12] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *ArXiv*, abs/2410.07864, 2024. 1

[13] Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. Cacti: A framework for scalable multi-task multi-scene visual imitation learning, 2023. 2

[14] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations, 2023. 2

[15] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019. 3

[16] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots, 2024. 2

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 4

[18] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17359–17369, 2025. 3

[19] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[20] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, page 245–254, New York, NY, USA, 1985. Association for Computing Machinery. 5

[21] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022. 3

[22] Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. Hrp: Human affordances for robotic pretraining. In *Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024. 3

[23] Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, and Hong Zhang. Functo: Function-centric one-shot imitation learning for tool manipulation. *ArXiv*, abs/2502.11744, 2025. 2, 3, 7

[24] Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, and Hong Zhang. Mimicfunc: Imitating tool manipulation from a single human video via functional correspondence. *arXiv preprint arXiv:2508.13534*, 2025. 3

[25] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[26] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *Robotics: Science and Systems*, 2025. 5

[27] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 6, 3

[28] TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, Naveen Kuppuswamy, Kuan-Hui Lee, Katherine Liu, Dale McConachie, Ian McMahon, Haruki Nishimura, Calder Phillips-Grafflin, Charles Richter, Paarth Shah, Krishnan Srinivasan, Blake Wulfe, Chen Xu, Mengchao Zhang, Alex Alspach, Maya Angeles, Kushal Arora, Vitor Campagnolo Guizilini, Alejandro Castro, Dian Chen, Ting-Sheng Chu, Sam Creasey, Sean Curtis, Richard Denitto, Emma Dixon, Eric Dusel, Matthew Ferreira, Aimee Goncalves, Grant Gould, Damrong Guoy, Swati Gupta, Xuchen Han, Kyle Hatch, Brendan Hathaway, Allison Henry, Hillel Hochsztein, Phoebe Horgan, Shun Iwase, Donovon Jackson, Siddharth Karamcheti, Sedrick Keh, Joseph Masterjohn, Jean Mercat, Patrick Miller, Paul Mitiguy, Tony Nguyen, Jeremy Nimmer, Yuki Noguchi, Reko Ong, Aykut Onol, Owen Pfannenstiehl, Richard Poyner, Leticia Priebe Mendes Rocha, Gordon Richardson, Christopher Rodriguez, Derick Seale, Michael Sherman, Mariah Smith-Jones, David Tago, Pavel Tokmakov, Matthew Tran, Basile Van Hoorick, Igor Vasiljevic, Sergey Zakharov, Mark Zolotas, Rares Ambrus, Kerri Fetzer-Borelli, Benjamin Burchfiel, Hadas Kress-Gazit, Siyuan Feng, Stacie Ford, and Russ Tedrake. A careful examination of large behavior models for multitask dexterous manipulation, 2025. 1

[29] Chen Wang, Linxi (Jim) Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning*, 2023. 1

[30] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models, 2024. 2

[31] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, 2024. 2

[32] Shijie Wu, Yihang Zhu, Yunao Huang, Kaizhen Zhu, Jiayuan Gu, Jingyi Yu, Ye Shi, and Jingya Wang. Afforddp: Generalizable diffusion policy with transferable affordance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6971–6980, 2025. 3

[33] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. *ArXiv*, abs/2502.16932, 2025. 1, 2, 5, 6

[34] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Dee M, Jodilyn Peralta, Brian Ichter, Karol Hausman, and Fei Xia. Scaling robot learning with semantically imagined experience, 2023. 2

[35] Zhecheng Yuan, Sizhe Yang, Pu Hua, Cancer Suk Chul Chang, Kaizhe Hu, Xiaolong Wang, and Huazhe Xu. Rl-vigen: A reinforcement learning benchmark for visual generalization. *ArXiv*, abs/2307.10224, 2023. 1

[36] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations, 2024. 1, 4

[37] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023. 3

[38] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024. 4, 6, 3

[39] Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. *arXiv preprint arXiv:2412.05268*, 2024. 2, 3, 4