This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AffordDP: Generalizable Diffusion Policy with Transferable Affordance

Shijie Wu† Yihang Zhu† Yunao Huang Kaizhen Zhu Jiayuan Gu Jingyi Yu
Ye Shi* Jingya Wang*
ShanghaiTech University, Shanghai, China
{wushj12023,zhuyh2023,huangya2024,zhukzh2024,gujy1,yujingyi,shiye,wangjingya}@shanghaitech.edu.cn

https://afforddp.github.io/

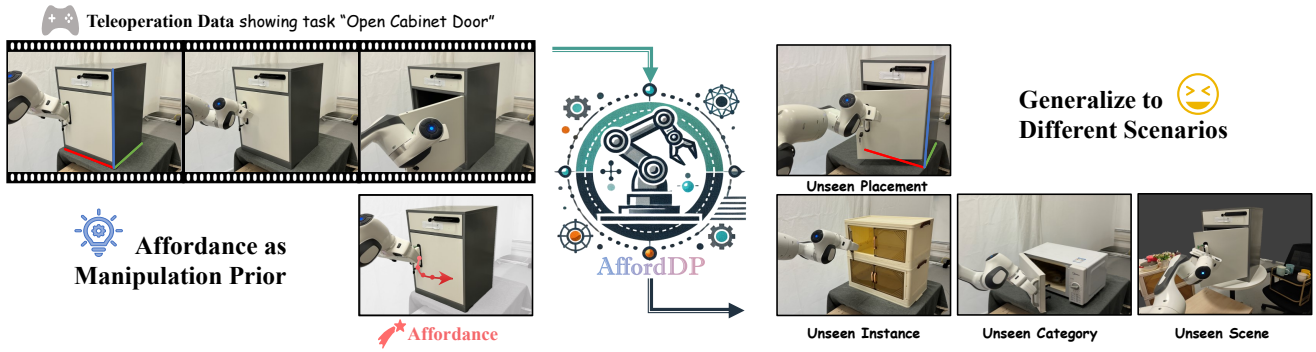

Figure 1. We propose AffordDP, a novel diffusion-based imitation learning method for generalizable robotic manipulation. AffordDP leverages rich manipulation priors from transferable affordances to enhance generalization to unseen scenarios and incorporates affordance guidance to enable precise control.

Abstract

Diffusion-based policies have shown impressive performance in robotic manipulation tasks while struggling with out-of-domain distributions. Recent efforts attempted to enhance generalization by improving the visual feature encoding for diffusion policy. However, their generalization is typically limited to the same category with similar appearances. Our key insight is that leveraging affordances—manipulation priors that define "where" and "how" an agent interacts with an object—can substantially enhance generalization to entirely unseen object instances and categories. We introduce the Diffusion Policy with transferable Affordance (AffordDP), designed for generalizable manipulation across novel categories. AffordDP models affordances through 3D contact points and post-contact trajectories, capturing the essential static and dynamic information for complex tasks. The transferable affordance from in-domain data to unseen objects is achieved by estimating a 6D transformation matrix using foundational vision models and point cloud registration techniques. More importantly, we incorporate affordance guidance during diffusion sampling that can refine action sequence generation. This guidance directs the generated action to gradually move towards the desired manipulation for unseen objects while keeping the generated action within the manifold of action space. Experimental results from both simulated and real-world environments demonstrate that AffordDP consistently outperforms previous diffusion-based methods, successfully generalizing to unseen instances and categories where others fail.

1. Introduction

Imitation learning offers an efficient solution for teaching robots a variety of skills by replicating behavior from experts without the need to explicitly program actions or model the environment. It has garnered significant interest across a wide range of fields, including autonomous driving [13, 16, 21], robotic manipulation [3, 6, 15, 17, 22, 34, 36], and legged locomotion [20, 25, 26]. Recently, diffusion-based imitation models [6, 14, 30] have gained significant attention within the field. These models conceptualize a robot's visuomotor policy as a conditional diffusion denoising process, facilitating the learning of conditional action distributions.

†Equal contributions.
*Corresponding authors.

Despite their high expressiveness and flexibility, diffusion-based imitation models [6, 14, 30] still struggle to generalize to out-of-domain distributions, which significantly limits their practical deployment. Recently, some pioneering works have sought to enhance the generalization ability of visual feature encoding and directly apply these advancements to diffusion policy [6]. For instance, Ze et al. [41] utilized 3D point cloud representations instead of 2D representations to improve feature encoding. Concurrently, Yang et al. [38] proposed integrating SIM(3)-equivariant networks to enhance generalization ability. Additionally, Wang et al. [35] introduced 3D semantic fields with multi-view input into diffusion policy to further improve its generalization ability. While these approaches may enable diffusion models to generalize to different initial locations, appearances, or poses, their generalization is typically limited to the same category with similar appearances. Subsequently, achieving the generalization of diffusion models to unseen instances and unseen categories remains challenging.

As human beings, we, in contrast, can easily transfer skills to objects with different appearances or even different types of objects. For example, once we learn the skill of opening a cabinet, it is easy to apply this skill to opening a microwave or a refrigerator (just as learning to ride a bicycle makes it easy to ride a tricycle). This ability is due to the vast amount of task-specific interaction prior knowledge that humans possess. This interaction prior knowledge at the object level specifically manifests as knowing "where" and "how" an agent should interact with an object to complete a given task, which is often defined as affordances in many studies [1, 40]. However, current imitation learning methods typically define affordances as 2D contact points with trajectory, task-specific keyframes, or value maps [2, 28, 42]. This significantly limits the representation of affordances in terms of "how" an agent should interact with the object. To tackle this key issue, we design a general representation for both static and dynamic affordances and further propose a diffusion policy with affordance transfer, allowing for the transfer of interaction prior knowledge across different instances and categories.

Specifically, we present a new Diffusion Policy with transferable Affordance that we call AffordDP to address complex manipulations of unseen objects and categories. The transferable affordance encompasses both static and dynamic affordances, capturing 3D contact points and post-contact trajectories, thus providing the "where" and "how" information crucial for manipulation tasks. The transferable affordance is derived from in-domain data and applied to unseen objects by estimating a 6D transformation matrix, utilizing foundational vision models and point cloud registration techniques. AffordDP employs the transferable affordance as an informative condition for the diffusion model, thereby enhancing its capability to generalize across varied manipulative scenarios. However, simply integrating transferable affordance as the diffusion condition may lack sufficient accuracy to achieve zero-shot adaptation for unseen robotic manipulation. To address this challenge, we introduce an affordance-guided sampling process for diffusion policy. In particular, the affordance guidance is defined as the distance between the robot's end-effector position and the static affordance. This guidance during the diffusion sampling process enables the generated action sequence to gradually move toward the desired manipulation for unseen objects while keeping the generated action within the manifold of action space. We adaptively apply the affordance guidance for fine-grained control. The novelty and contributions of this paper are outlined as follows.

- We introduce AffordDP that leverages transferable affordance for enhanced generalization to unseen objects and categories that cannot be completed by existing diffusion-based methods.
- Our framework enables the transfer of both static and dynamic affordances by utilizing foundational vision models, making it possible to represent complex manipulations with limited in-domain data.
- We integrate adaptive affordance guidance to our AffordDP that can direct the generated action sequence to gradually move towards the desired manipulation for unseen objects within the manifold of action space.

2. Related work

2.1. Diffusion-based Imitation Learning

As one of the first works [6, 14, 30] that introduce diffusion into imitation learning, the Diffusion Policy [6] leverages a diffusion denoising process to model conditional action distributions. Despite its advantages in modeling multimodal data and providing stable training, the Diffusion Policy still faces limitations in generalizing to out-of-domain distributions. When the position or shape of objects changes, the Diffusion Policy struggles to adapt successfully. 3D Diffusion Policy [41] addresses this to some extent by using point clouds for visual representation, which improves spatial generalization. Equibot [38] further advances this by combining SIM(3)-equivariant neural network architectures with diffusion policies, enhancing generalization to SIM(3)-equivariant scenes. Another approach, Im2Flow2Act [37], combines self-supervised 2D flow prediction with a flow-based Diffusion Policy. However, its reliance on 2D flow limits its effectiveness, failing to capture object movements in the 3D space accurately and leading to less precise actions. Concurrent with our work, GenDP [35] and G3Flow [4] employ 3D semantic fields to enhance point cloud observations used in Diffusion Policy. However, these approaches require a multi-view setup to generate semantic

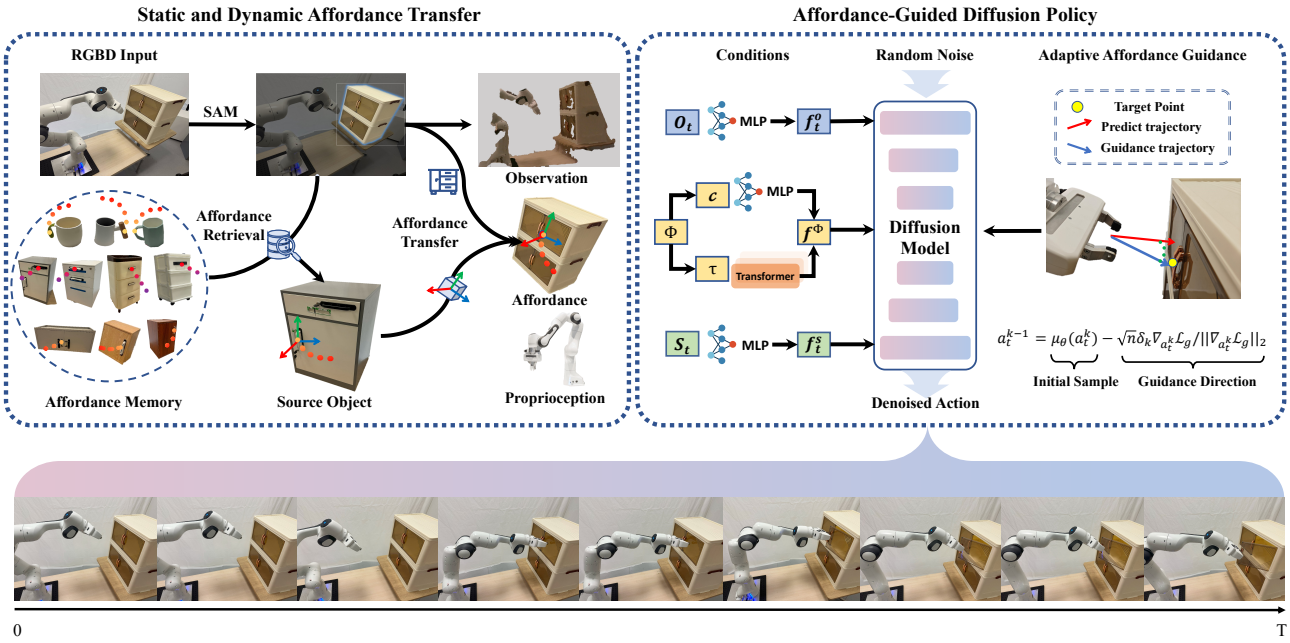

Figure 2. Overview of AffordDP. The left part demonstrates static and dynamic affordance transfer. Given the target scene RGB-D image, AffordDP retrieves a similar object in the affordance memory and transfers its static and dynamic affordance to the target object. The right part illustrates the key components of affordance-guided diffusion policy. Conditioned on 3D affordance, point cloud observation, and robot proprioception, AffordDP utilizes the Diffusion Policy and adaptive affordance guidance for precise control.

fields and are limited to category-level generalization. In contrast, our method uniquely conditions diffusion on 3D affordance trajectories, allowing our model to generalize more effectively to objects with similar affordance trajectories, regardless of their category.

2.2. Generalizable Manipulation for Robotics

Generalizable manipulation is a principal focus in the field of imitation learning. To encourage policies to concentrate on task-relevant factors while minimizing irrelevant distractions, previous methods typically extract object-centric features from images, like object poses [8], task-specific region proposals [9], keypoints [23] and segmented point clouds [46]. A further trend in this field is the integration of affordance into policy learning. Mainstream works [2, 12, 24] tend to learn affordance from large datasets or through intensive reinforcement learning, which can entail significant overhead. To address these limitations, recent advancements have leveraged semantic information from foundational vision models to enable zero-shot affordance transfer. Pioneering methods such as Robo-ABC [18] and RAM [19] allow affordance transfer in a training-free manner. However, they are limited to predicting affordances either as 2D contact points or 3D contact points with direction. Additionally, they rely on grasp generators [10] and motion planners [7, 32] to perform manipulation tasks, which face challenges in high-dimensional action spaces and complex manipulation tasks. This reliance often results in unimodal actions, limiting effectiveness in environments with significant uncertainty. Our method advances these developments by combining 2D semantic correspondences with 3D geometrical correspondences, significantly enhancing the capability for effective trajectory-level transfer.

3. Transfer of Static and Dynamic Affordances

AffordDP generates 3D affordances denoted by $\Phi = (c, \tau)$, where $c \in \mathbb{R}^3$ represents the static affordance as 3D contact point and $\tau \in \mathbb{R}^{3\times N}$ represents the dynamic affordance as post-contact trajectory in the form of an ordered list of 3D points. Affordance transfer is defined as generating the affordance for the target object given a source object with its affordance. Upon specifying a target object, AffordDP retrieves the most similar object from the object memory. The static affordance c is transferred using the semantic correspondence, and the dynamic affordance τ is transferred using a 6D transformation matrix, whose translation and rotation components are derived independently. An overview is shown in Fig. 3.

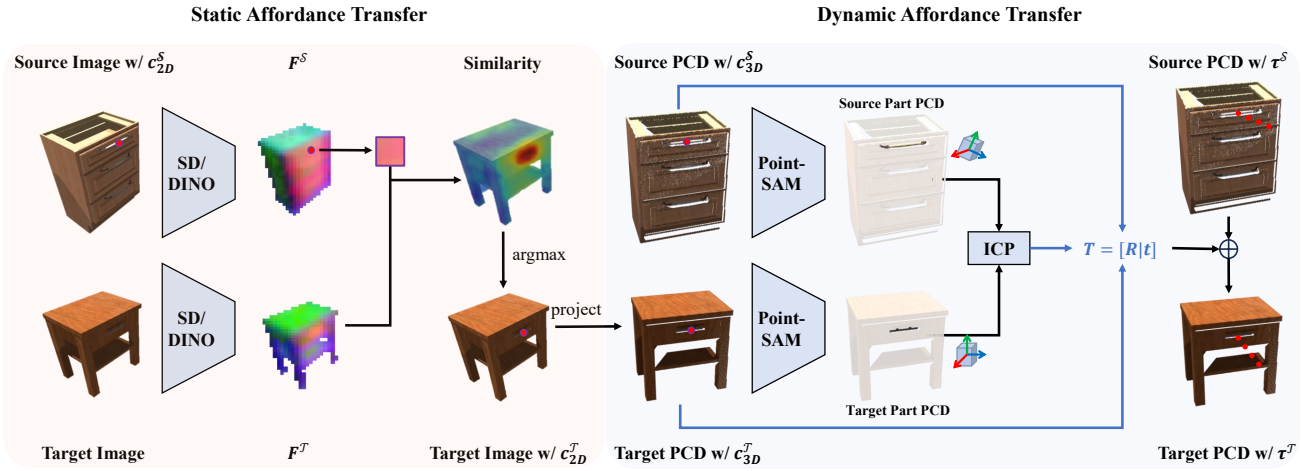

Figure 3. Overview of Static and Dynamic Affordance Transfer. The left part is static affordance transfer, and the right part is dynamic affordance transfer. The source and target images are processed through SD-DINOv2 [43] to generate feature maps F^S and F^T. The similarity is computed and used to find the corresponding target image points using an argmax operation. These points are projected back to obtain point clouds. Point-SAM [45] is then used to obtain the source and target part point clouds. The Iterative Closest Point (ICP) algorithm is applied to the point clouds to determine the transformation matrix $T = [\mathbf{R}|\mathbf{t}]$ for dynamic affordance transfer.

3.1. Affordance memory and similarity retrieval

To build the affordance memory, we utilize Grounded-SAM [29] to crop the object from the image and employ the CLIP [27] image encoder to extract the feature vector z. Additionally, we record the contact point as the static affordance c, and the trajectory of the robot end-effector after contacting the object (in world coordinates) as the dynamic affordance τ. We define the affordance memory $\mathcal{M}$, which includes the task name T, the affordance $\Phi = \{c, \tau\}$, the appearance feature z, and the point cloud of the manipulable object $\mathcal{P}$ recovered from RGB-D images as follows,

$$\mathcal{M} = \{(T, \Phi, z, \mathcal{P})\}. \tag{1}$$

The cosine similarity between the source and target cropped image feature vectors serves as a proximity metric for retrieval. In practical terms, for an object with task T, we perform retrieval within the subset of the memory corresponding to the same task T.

3.2. Static affordance transfer

Static affordance transfer involves finding the contact point of the target object in 3D world coordinates, given the contact point of the source object. This task is challenging because the two objects may differ in appearance, shape, rotation, or even category. We leverage the rich semantic information provided by foundation vision models to mitigate these variances. We upsample SD-DINOv2 [43] features to generate pixel-level image features. By matching the features of the source static affordance in the 2D image $c_{2D}^{\mathcal{S}}$ with all features of the target image pixels, we can identify the target static affordance in the 2D image $c_{2D}^{\mathcal{T}}$. We then back-project the target static affordance $c_{2D}^{\mathcal{T}}$ to 3D world coordinates as $c_{3D}^{\mathcal{T}}$.

3.3. Dynamic affordance transfer

To transfer dynamic affordances, we utilize a 6D transformation matrix. Previous alignment-based methods that rely on instance-level registration between source and target objects [47] often yield unreliable results due to shape and scale discrepancies. For example, transferring a top drawer affordance from a tall closet to a central drawer in a square closet can lead to inaccuracies. To address this issue, we independently estimate the rotation matrix and translation vector, better accommodating structural differences and enhancing the accuracy of affordance transfer.

The translation vector $\mathbf{t}$ is obtained by computing the difference between the source and target 3D contact points. To determine the rotation matrix $\mathbf{R}$, we perform point cloud registration on source and target point clouds. We employ a state-of-the-art promptable 3D segmentation model [45], using c_{3D} as a prompt to segment manipulable parts. This approach reduces sensitivity to structural variance, as part point clouds exhibit less variability than complete objects. We then use the iterative closest point (ICP) algorithm [5] to register the 3D parts by minimizing the objective function:

$$E(\mathbf{T}) = \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}} \left[(\mathbf{p} - \mathbf{T}\mathbf{q}) \cdot \mathbf{n}_{\mathbf{p}}\right]^2, \tag{2}$$

where $\mathbf{T}$ is the transformation, $\mathcal{K} = \{(\mathbf{p}, \mathbf{q})\}$ is the correspondence set between the target point cloud $\mathbf{P}$ and the

source point cloud $\mathbf{Q}$, and $\mathbf{n_p}$ is the normal of point $\mathbf{p}$[*].

After deriving the rotation matrix $\mathbf{R}$ from the registration results, we combine it with the translation vector $\mathbf{t}$ obtained earlier to compute the 6D transformation matrix:

$$\tau^{\mathcal{T}} = \mathbf{R}\tau^{\mathcal{S}} + \mathbf{t}. \tag{3}$$

This 6D transformation matrix maps the dynamic affordance of the source object to the target object.

4. Affordance Guided Diffusion Policy

To enhance the generalization capacity, we design a conditional diffusion model that takes in the transferable affordance as an additional condition, which treats this manipulation prior as a bridge to narrow the gap between training and practical deployment. However, simply adding affordance as a diffusion condition does not fully leverage its potential, as the diffusion model is inherently focused on modeling the overall action distribution, prioritizing alignment with the distribution rather than adherence to specific task constraints. To address this, we introduce an adaptive affordance-guided sampling process to direct the policy toward static affordance (contact point), enhancing task focus and precision.

4.1. Conditions for AffordDP

The overall conditions of our diffusion model are:

$$\mathcal{C} = \{O_t, S_t, \Phi\}, \tag{4}$$

where O_t represents the scene point cloud, S_t denotes robot proprioception, and Φ represents the affordance, which is composed of static affordance c (contact point) and dynamic affordance τ (trajectory). Our diffusion model builds upon Denoising Diffusion Implicit Models (DDIMs) [31] and aims to approximate affordance-guided conditional distribution $p(a_t|O_t, S_t, \Phi)$, to capture the action distribution through a gradual denoising process from a Gaussian distribution.

To extract the overall conditioning feature f_t from each element in $\mathcal{C}$, we use distinct encoders. Since only a small region surrounding the robot and target object offers meaningful contextual information, we crop and downsample the scene point cloud before applying an MLP for scene feature extraction, following a similar approach to the 3D Diffusion Policy [41]. To capture the full dynamics of the trajectory sequence τ, we utilize a transformer encoder [33] with a [CLS] token to represent the trajectory feature. The contact point c and robot proprioception S_t features are extracted using separate MLPs. We define the conditioning feature f_t as follows,

$$f_t = \{f_t^O, f_t^S, f^\Phi\}, \tag{5}$$

where f_t^O denotes the scene feature, f_t^S represents the robot state feature, and f^Φ integrates both the static and dynamic affordance features for simplicity.

4.2. Conditional diffusion model training

Starting from an initial noise vector a_t^k, AffordDP uses a noise prediction network ϵ_θ to iteratively denoise the noise vector conditioned on the conditioning feature f_t till a noise-free action a_t^0 is formed. This process can be described using the following equations:

$$\begin{aligned} a_t^{k-1} =& \sqrt{\alpha_{k-1}}(a_t^k - \sqrt{1-\alpha_k}\epsilon_\theta^k)/\sqrt{\alpha_k} + \\ & \sqrt{1-\alpha_{k-1}-\sigma_k^2}\epsilon_\theta^k + \sigma_k \epsilon^k, \end{aligned} \tag{6}$$

$$\sigma_k = \sqrt{(1-\alpha_{k-1})/(1-\alpha_k)}\sqrt{1-\alpha_k/\alpha_{k-1}}\,, \tag{7}$$

where ϵ^k is random noise, α_k is the noise schedule and k is the number of denoising iterations. Similar to previous diffusion-based policies [6, 41], we train the noise prediction network ϵ_θ to predict the noise ϵ^k added at k-th iteration,

$$\mathcal{L} = ||\epsilon_k - \epsilon_\theta(a_t^k, k, f_t)||^2. \tag{8}$$

4.3. Diffusion sampling with affordance guidance

However, while the standard diffusion formulation is designed to capture the comprehensive action distribution, it may generate actions that remain within this distribution but fail to meet specific task conditions, particularly in high-precision tasks such as grasping a door handle to open it. To this end, we add additional guidance to steer its output distribution towards specific properties y. With the Bayes Theorem, we can introduce the given condition with an additional likelihood term $p(a_t|y)$:

$$\nabla_{a_t^k} \log p(a_t^k|y) = \nabla_{a_t^k} \log p(a_t^k) + \nabla_{a_t^k} \log p(y|a_t^k). \tag{9}$$

By utilizing a differentiable loss function $L(a_t^0, y)$ for sampling guidance, we eliminate the need to train the classifier or retrain the whole network. We can use the estimated likelihood of $L(a_t^0, y)$ for the additional correction step:

$$a_t^{k-1} = \text{DDIM}(a_t^k, \boldsymbol{\epsilon}_\theta(a_t^k, k, f_t)) - \gamma \nabla_{a_t^k} L(\hat{a}_t^0(a_t^k), y)\,, \tag{10}$$

where $\hat{a}_t^0(a_t^k)$ is an estimate based on a_t^k and γ is guidance weight.

During inference, given the robot's current joint positions q_{pos}, we use the forward kinematics to compute the 3D position of the robot's end-effector p_{ee}. We then use the distance between the end-effector's position p_{ee} and the static affordance c_{3D} as the adaptive loss function $\mathcal{L}_g$ for the adaptive affordance guidance. This loss function is adaptive because it is only utilized when the end-effector approaches

[*]We directly apply the implementation of Open3D [44], correspondence set $\mathcal{K}$ and normal $\mathbf{n_p}$ are also estimated by Open3D.

the static affordance c_{3D} within a predefined range. This prevents it from being applied when the gripper is far from or has already grasped the target object. If the gripper is approaching the target object, then the adaptive loss function can be represented as:

$$\mathcal{L}_g = \begin{cases} \|p_{ee} - c_{3D}\|_2, & \|p_{ee} - c_{3D}\|_2 < \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Here θ is a predefined threshold to control the actuating range of the affordance guidance. Motivated by DSG [39], we introduce an analytical solution to the sampling process of the diffusion model, which can guide the generation process to the direction of gradient descent while preserving the generated a^0 within the manifold:

$$a_t^{k-1^*} = \mu_\theta(a_t^k) - \sqrt{n}\delta_k \nabla_{a_t^k} \mathcal{L}_g / ||\nabla_{a_t^k} \mathcal{L}_g||_2, \tag{12}$$

$$\begin{aligned} \mu_\theta(a_t^k) = & \sqrt{\alpha_{k-1}}(a_t^k - \sqrt{1-\alpha_k}\epsilon_\theta)/\sqrt{\alpha_k} + \\ & \sqrt{1-\alpha_{k-1}-\sigma_k^2}\epsilon_\theta \,, \end{aligned} \tag{13}$$

where n represents the action dimensions. By incorporating affordance loss guidance into the standard diffusion formulation, AffordDP can generate action sequences directed toward the target manipulation positions for unseen objects, thereby enhancing its generalization capabilities.

5. Experiments

5.1. Experiment Setup

We evaluate our approach on multiple tasks, including both simulations and real-world scenarios. For the simulation scenario, we use IsaacGym as the simulation platform, leveraging objects from GAPartnet [11] and point clouds generated by depth cameras within the simulation for observations. We use the Franka arm and a side-mounted camera setting for both simulation and real-world scenarios.

Expert demonstrations. We use CuRoBo [32] as the motion planner for expert demonstration collection in the simulation. To ensure diversity in the collected trajectories, we introduce random noise to the initialization parameters, including the 6D object position, the robot arm's base position, and the initial DOF of the robot arm. Using part bounding boxes annotated by GAPartnet, we can collect expert demonstrations specific to part manipulation. In the real-world scenario, we collect expert demonstrations through teleoperation.

Baseline methods. We selected two state-of-the-art diffusion-based policies as our baselines for comparison: 1) Diffusion Policy (DP) [6], which utilizes images as observations to generate actions via diffusion; and 2) DP3 [41], which replaces image inputs in the Diffusion Policy with point clouds. We evaluate different policies by the success rate.

Training settings. DP [6] and DP3 [41] adopt a single-object training approach, involving only one object during the training process. This setup encourages the policy to learn object-specific features, potentially limiting its generalization ability to adapt to unseen instances, which can be a common requirement in practical applications. To this end, we design two training settings: object-specific policy training and unified policy training. The former retains the original single-object training setup used in DP and DP3, while the latter employs a multi-object training approach that promotes learning task-relevant strategies for broader generalization.

5.2. Simulation Performance

We evaluate our method on two tasks for the simulation scenario: PullDrawer and OpenDoor. For a fair comparison, we assess all methods under both training settings—object-specific policy training and unified policy training—to highlight our method's consistent advantage over baselines.

In the object-specific policy training, we collect 30 expert demonstrations for each task. To analyze how variability in training data impacts algorithm performance, we construct datasets with differing levels of variance, reflected in the initial 6D placements of the objects and the initial joint positions of the robot arm. Our method and the baselines are evaluated across each variance level to assess robustness. The quantitative results are summarized in Tab. 2. Our approach significantly outperforms the baselines across different variance levels on both tasks. Notably, while baseline methods are comparable to ours under low variance conditions, our method demonstrates a substantial advantage over baselines as variance levels increase.

For the unified policy training, we carefully choose 5 distinct objects for each task and collect 20 demonstrations for each object. To evaluate generalization capabilities, we introduce three types of test objects: seen instances, unseen instances within the same category, and instances from unseen categories. The quantitative results presented in Tab. 3 show that our approach achieves superior performance across all three object types. Even for some categories that are not involved during training, our method still demonstrates strong generalization capabilities and exceptional performance. In contrast, the baseline methods exhibit very low success rates with unseen categories. By leveraging transferable affordance, AffordDP can capture the inherent properties of both seen and unseen objects, significantly enhancing the policy's learning and adaptability.

5.3. Real-world Performance

We evaluate our method on three tasks for the real-world scenario: OpenDoor, PullDrawer, and Pick&Place. We use the unified policy training setting for real-world experiments. Using teleoperation, we collect 25 demonstrations

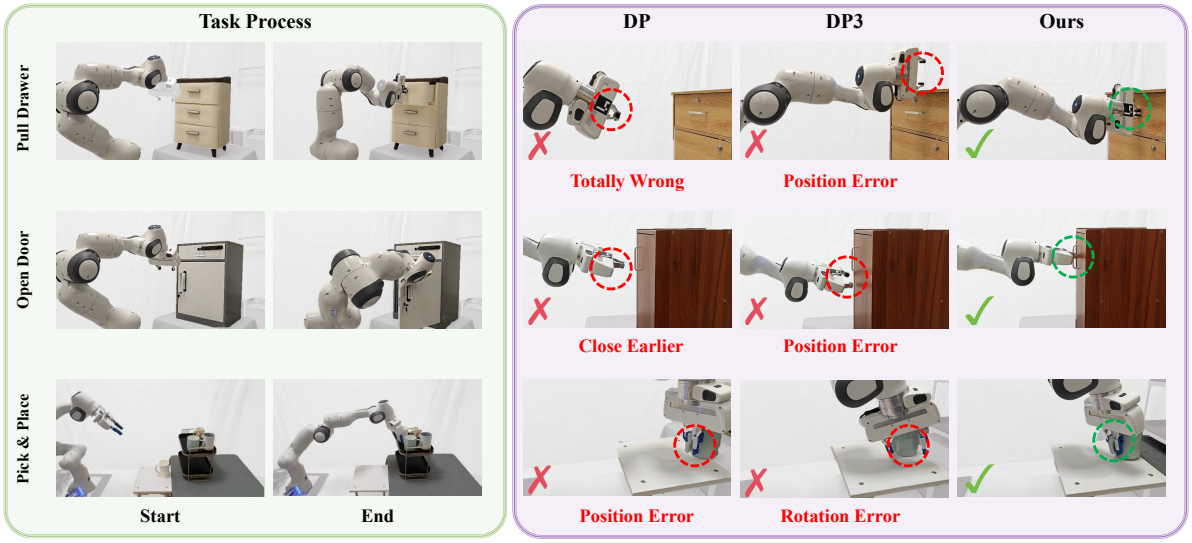

Figure 4. Policy rollouts in the real world. The left part is an intuitive demonstration of three real-world task processes, involving PullDrawer, OpenDoor, and Pick&Place. The right part represents the policy rollout results. Without spatial perception and a lack of adequate training data, which can be over 200 demonstrations [6, 41], DP fails to target the object correctly, often resulting in random and potentially unsafe actions. Lacking comprehensive static and dynamic affordance information, DP3 struggles to grasp the target accurately, even if it moves the gripper close to the target. In contrast, our method effectively grasps the target object and completes the task, demonstrating accurate spatial targeting.

Inst. & Cate. Generalization	Cabinet 1	Cabinet 2	Dressing Table	Oven
DP [6]	✗	✗	✗	✗
DP3 [41]	✗	✗	✗	✗
Ours	✓	✓	✓	✓

Table 1. Instance and Category Generalization on PullDrawer task. We evaluate the generalization capacities by replacing the target object with unseen instances and unseen categories. Each object undergoes three evaluation trials, and a green checkmark is awarded if the task succeeds in at least one out of the three trials.

per object, with a total of 4 objects for each task, resulting in 100 demonstrations per task.

Consistent with the unified policy training experiments for the simulation scenario, we evaluate our method on three types of test objects: seen instances, unseen instances within the same category, and instances from unseen categories. The results, shown in Tab. 4, highlight the effectiveness of our method in real-world scenarios. Fig. 4 shows policy deployment results, showcasing our methods's supe-

Simulation results of object-specific policy for seen instance						
Tasks	Pull Drawer			Open Door		
Methods	Easy	Median	Hard	Easy	Median	Hard
DP	**80.0%**	16.7%	13.3%	76.7%	26.7%	30.0%
DP3	**80.0%**	13.3%	16.7%	66.7%	36.7%	20.0%
Ours	**80.0%**	**33.3%**	**26.7%**	**90.0%**	**53.3%**	**50.0%**

Table 2. Simulation performance of different methods under the object-specific policy training setting for seen instance. We categorized datasets into varying difficulty levels based on variance levels. The variance level is directly related to the random noise (position, rotation) introduced during the data collection process, the details of variance levels can be found in supplementary.

Simulation results of the unified policy training setting						
Tasks	Pull Drawer			Open Door		
Methods	Seen Instance	Unseen Instance	Unseen Category	Seen Instance	Unseen Instance	Unseen Category
DP	20.7%	6.7%	3.3%	23.3%	17.5%	3.3%
DP3	41.3%	10.0%	3.3%	41.1%	5.0%	5.6%
Ours	**90.0%**	**55.6%**	**73.3%**	**74.4%**	**52.5%**	**26.7%**

Table 3. Simulation performance of different methods under the unified policy training setting. We compare our method with the baseline across three types of test objects, involving seen objects, unseen instances within the same category, and instances from unseen categories.

Real-World Results									
Tasks	Pull Drawer			Open Door			Pick & Place		
Methods	Seen Instance	Unseen Instance	Unseen Category	Seen Instance	Unseen Instance	Unseen Category	Seen Instance	Unseen Instance	Unseen Category
DP	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
DP3	10.0%	0.0%	10.0%	20.0%	0.0%	30.0%	40.0%	30.0%	15.0%
Ours	**67.5%**	**45.0%**	**40.0%**	**80.0%**	**50.0%**	**50.0%**	**52.5%**	**65.0%**	**50.0%**

Table 4. Real-World performance of different methods for all three types of test objects for three tasks. Our method consistently outperforms DP and DP3 in all scenarios, especially in tasks with unseen instances and categories. Notably, DP fails in all tasks, while DP3 achieves limited success in certain unseen categories. DP may require over 200 demonstrations for training before showing reliable performance [6, 41]. Observations from policy rollouts indicate that DP3 has overfitted to specific positions, allowing successful grasps only when the test object is located in those specific areas. Additionally, DP3 often produces highly overlapping action sequences across different object placements in task scenarios involving unseen instances and categories. In contrast, our method demonstrates a stable superiority. We can conclude from the task results of the seen instance that static and dynamic affordance provides a bridge across action distributions associated with different objects—something DP3 struggles to model due to the complexity of such distributions.

riority over baselines. And Tab. 1 gives some results on the PullDrawer task to showcase the instance and category generalization capacity of our method. Experimental outcomes demonstrate that our method successfully generalizes to unseen instances, even those from unseen categories. In the most challenging scenarios, our method continues to exhibit strong zero-shot capabilities, significantly outperforming baseline approaches.

Contact Point	Trajoctary	Affordance Guidance	Seen Instance	Unseen Instance	Unseen Category
✓			72.2%	42.5%	17.8%
✓	✓		71.6%	50.0%	22.2%
✓	✓	✓	**74.4%**	**52.5%**	**26.7%**

Table 5. Ablation study in simulation. Success rates are shown across three scenarios: seen objects, unseen instances, and unseen categories. The combination of contact point, trajectory, and affordance guidance yields the highest success rates in both seen and unseen scenarios, particularly excelling in unseen categories.

5.4. Ablation Study

To validate the efficacy of our proposed technique, we conducted ablation studies in a simulated environment under the unified policy training setting. In these experiments, we selectively omitted two critical components, trajectory and affordance guidance, in our method while maintaining other modules unchanged, to evaluate the importance of these components.

As illustrated in Tab. 5, the performance of AffordDP markedly decreases when only the contact point is included, and the dynamic affordance is excluded. The integration of dynamic affordance leverages prior knowledge of object manipulation, thereby enhancing the model's performance, particularly for tasks with significant variability in object manipulation trajectories.

Moreover, adaptive affordance guidance substantially boosts the performance of AffordDP in previously unseen categories. When the model encounters objects with appearances that differ considerably from those in the training set, adaptive affordance guidance steers the generated action to gradually move toward the desired manipulation for unseen objects while keeping the generated action within the manifold of action space, effectively mitigating the impact of covariance shift.

6. Conclusion and Discussion

In conclusion, the Generalizable Diffusion Policy with transferable Affordance (AffordDP) significantly advances the generalization capabilities of diffusion-based robotic manipulation policies. By modeling affordances through 3D contact points and post-contact trajectories, and incorporating affordance guidance during diffusion sampling, AffordDP effectively generalizes across entirely unseen object instances and categories. The estimation of the 6D transformation matrix using foundational vision models and point cloud registration techniques allows for the transferability of these affordances. Experimental results in both simulated and real-world environments validate the superiority of AffordDP over previous methods, marking a substantial step forward in the robotic community.

Despite these advancements, AffordDP does have some limitations that need to be addressed in future work. Current foundational models have limitations in understanding spatial information, which can restrict AffordDP's ability to specify affordance transfers for particular parts. Additionally, AffordDP's effectiveness is limited in scenarios where modeling or extracting affordances is challenging. Future work should focus on enhancing the spatial understanding of foundational models and improving affordance extraction techniques to overcome these limitations.

7. Acknowledgement

This work was supported by the Shanghai Local College Capacity Building Program (23010503100), National Natural Science Foundation of China (No.62406195), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), HPC Platform of ShanghaiTech University and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

[1] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020. 2

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. 2, 3

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1

[4] Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, et al. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation. *arXiv preprint arXiv:2411.18369*, 2024. 2

[5] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 4

[6] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. 1, 2, 5, 6, 7, 8

[7] David Coleman, Ioan Sucan, Sachin Chitta, and Nikolaus Correll. Reducing the barrier to entry of complex robotic software: a moveit! case study. *arXiv preprint arXiv:1404.3785*, 2014. 3

[8] Xinke Deng, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox. Self-supervised 6d object pose estimation for robot manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3665–3671. IEEE, 2020. 3

[9] Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep object-centric representations for generalizable robot learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7111–7118. IEEE, 2018. 3

[10] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 2023. 3

[11] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 6

[12] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023. 3

[13] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 1

[14] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *Conference on Robot Learning*, pages 3766–3777. PMLR, 2023. 1, 2

[15] Siddhant Haldar, Jyothish Pari, Anant Rai, and Lerrel Pinto. Teach a robot to fish: Versatile imitation from one minute of demonstrations. *arXiv preprint arXiv:2303.01497*, 2023. 1

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[17] Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilue Wang, Longzan Luo, Lily Lee, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, et al. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. *arXiv preprint arXiv:2411.18623*, 2024. 1

[18] 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. *arXiv preprint arXiv:2401.07487*, 2024. 3

[19] 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

[20] Tingguang Li, Yizheng Zhang, Chong Zhang, Qingxu Zhu, Jiapeng Sheng, Wanchao Chi, Cheng Zhou, and Lei Han. Learning terrain-adaptive locomotion with agile behaviors by imitating animals. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 339–345. IEEE, 2023. 1

[21] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1

[22] Qi Lv, Hao Li, Xiang Deng, Rui Shao, Yinchuan Li, Jianye Hao, Longxiang Gao, Michael Yu Wang, and Liqiang Nie. Spatial-temporal graph diffusion policy with kinematic mod-

eling for bimanual robotic manipulation. *arXiv preprint arXiv:2503.10743*, 2025. 1

[23] 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

[24] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 3

[25] Xinlei Pan, Tingnan Zhang, Brian Ichter, Aleksandra Faust, Jie Tan, and Sehoon Ha. Zero-shot imitation learning from demonstrations for legged robot visual navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 679–685. IEEE, 2020. 1

[26] Xue Bin Peng, Erwin Coumans, Tingnan Zhang, Tsang-Wei Lee, Jie Tan, and Sergey Levine. Learning agile robotic locomotion skills by imitating animals. *arXiv preprint arXiv:2004.00784*, 2020. 1

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[28] Krishan Rana, Jad Abou-Chakra, Sourav Garg, Robert Lee, Ian Reid, and Niko Suenderhauf. Affordance-centric policy learning: Sample efficient and generalisable robot policy learning using affordance-centric task frames. *arXiv preprint arXiv:2410.12124*, 2024. 2

[29] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4

[30] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. *arXiv preprint arXiv:2304.02532*, 2023. 1, 2

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[32] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, et al. Curobo: Parallelized collision-free robot motion generation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8112–8119. IEEE, 2023. 3, 6

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 5

[34] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 1

[35] Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kelestemur, Jiuguang Wang, and Yunzhu Li. Gendp: 3d semantic fields for category-level generalizable diffusion policy. In *8th Annual Conference on Robot Learning*. 2

[36] Zhenyu Wu, Yuheng Zhou, Xiuwei Xu, Ziwei Wang, and Haibin Yan. Momanipvla: Transferring vision-language-action models for general mobile manipulation. *arXiv preprint arXiv:2503.13446*, 2025. 1

[37] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024. 2

[38] Jingyun Yang, Zi-ang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning. *arXiv preprint arXiv:2407.01479*, 2024. 2

[39] Lingxiao Yang, Shutong Ding, Yifan Cai, Jingyi Yu, Jingya Wang, and Ye Shi. Guidance with spherical gaussian constraint for conditional diffusion. In *International Conference on Machine Learning*, 2024. 6

[40] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. *arXiv preprint arXiv:2412.01550*, 2024. 2

[41] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems*, 2024. 2, 5, 6, 7, 8

[42] Yantian Zha, Siddhant Bhambri, and Lin Guan. Contrastively learning visual attention as affordance cues from demonstrations for robotic grasping. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7835–7842. IEEE, 2021. 2

[43] 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, 2024. 4

[44] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 5

[45] Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, Fanbo Xiang, and Hao Su. Point-sam: Promptable 3d segmentation model for point clouds. *arXiv preprint arXiv:2406.17741*, 2024. 4

[46] Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. *arXiv preprint arXiv:2310.14386*, 2023. 3

[47] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024. 4