

# GenDP: 3D Semantic Fields for Category-Level Generalizable Diffusion Policy

Anonymous Author(s)

Affiliation

Address

email

## 1 Contents

2	<b>1 Method Details</b>	<b>1</b>
3	<b>2 Experiments Details</b>	<b>2</b>
4	2.1 Setup . . . . .	2
5	2.2 Demonstrations . . . . .	2
6	2.3 Training . . . . .	3
7	2.4 Inference . . . . .	3
8	2.5 Evaluation . . . . .	3
9	<b>3 Additional Experiments</b>	<b>3</b>
10	3.1 Quantitative Experiment Table . . . . .	3
11	3.2 Extension to ACT . . . . .	4
12	3.3 Ablation on Feature Extraction Backbone . . . . .	4
13	3.4 More Baselines . . . . .	6
14	3.4.1 SORNet . . . . .	6
15	3.4.2 DINOBot . . . . .	7
16	3.4.3 Results . . . . .	7
17	3.5 t-SNE . . . . .	8
18	<b>4 Detailed Comparisons with Prior Works</b>	<b>8</b>

## 19 1 Method Details

20 We use PointNet++ to encode one latent vector from the point clouds [1]. There are several key  
21 details regarding its usage.

- 22 • No BatchNorm Layer: In the diffusion policy implementation, exponential moving aver-  
23 ages (EMA) of the model parameters are used during the testing, which is not compatible  
24 with BatchNorm.
- 25 • Point Set Abstraction: Point set abstraction is an important operation in PointNet++. In  
26 our network, we deploy three layers of point set abstraction. For the first set abstraction,

27 we have 512 groups with a radius of 0.04 and 32 points. For the second set abstraction,  
 28 we have 128 groups with a radius of 0.08 and 64 points. For the last set abstraction, all  
 29 sampled points are grouped to extract one latent vector.

## 30 2 Experiments Details

### 31 2.1 Setup

32 We summarize object categories and the number of object instances we considered in Table 1.

Task Category	Task Name	#Demo	Object	#Train Instances	#Test Instances
Sim.	Hang Mug	200	Mug	4	2
	Insert Pencil	100	Pencil	2	1
Geo. Details	Place Can	60	Soda Can	3	3
	Use Toothbrush	60	Toothbrush	1	4
Cat. Gen.	Align Shoes	60	Shoe	3	5
	Use Spoon	60	Spoon	2	2
Geo. Ambi.	Collect Knife	60	Knife	1	2
	Open Pen	60	Marker Pen	1	3

Table 1: **Task Details Summary.** This table summarizes our task categories, specific tasks, objects, the number of demonstrations, training instances, and testing instances. We evaluate our framework on eight tasks and eight object categories, where each object category includes several instances with diverse appearances and shapes. We divide our tasks into four categories to study our method’s performance under various task types and scenarios, such as simulation and the real world.

33 Here is a list of task descriptions:

- 34 • **Hang Mug (Simulation):** Hang a randomly placed mug on a fixed mug tree.
- 35 • **Pencil Insertion (Simulation):** Pick up a pencil from the table and insert it into the pencil  
 36 sharpener.
- 37 • **Collect Knife:** Collect a lying knife into an open container, and the knife direction is  
 38 randomly chosen.
- 39 • **Open Pen:** Bimanually grasp the pen and open it.
- 40 • **Flip Can:** Flip a lying soda can and place it upright.
- 41 • **Use Toothbrush:** Grasp the toothbrush and spread the toothpaste on it.
- 42 • **Align Shoes:** Push shoes towards left.
- 43 • **Use Spoon:** Grasp spoon and scoop materials.

44 Due to space limitations, in all tables, we will use the following abbreviation:

- 45 • **Sim.:** Simulation
- 46 • **Geo. Details:** Geometry Details
- 47 • **Geo. Ambi.:** Geometry Ambiguity
- 48 • **Cat. Gen.:** Category Generalization

### 49 2.2 Demonstrations

50 In the simulation, we use SAPIEN to create simulation environments. We build an oracle policy to  
 51 generate demonstrations in the simulation. The generated dataset contains robot states, multi-view  
 52 RGBD observations, and robot actions.

53 In the real world, the data collection pipeline is similar, with additional calibration steps. We define  
54 the ChArUco board as the world coordinate and calibrate the camera poses to this coordinate. Then,  
55 we manually calibrate the transformation between the robot and the world coordinate. After the  
56 calibration phase, the data collection pipeline records multi-view RGBD observations, robot states,  
57 and robot actions.

## 58 **2.3 Training**

59 Before training, we preprocess multi-view RGBD observations into 3D semantic fields using the  
60 frozen DINOv2 encoder and save them to CPU memory. During training, the 3D semantic fields  
61 and corresponding ground truth actions are randomly sampled and fed into the diffusion policy.  
62 The training of the diffusion policy follows the practice outlined in the original Diffusion Policy  
63 paper [2].

## 64 **2.4 Inference**

65 During inference, the diffusion policy takes in one frame of multi-view RGBD observation and  
66 converts it into 3D semantic fields. Then, the 3D semantic fields are input into the diffusion policy  
67 to predict an action sequence.

68 We measure our inference time on the knife-collecting task, repeating the inference steps 10 times  
69 and computing the average time. Each inference step takes 0.226 seconds. The predicted action  
70 sequence is at 10 Hz, meaning the time difference between two consecutive actions in the action se-  
71 quence is 0.1 seconds. This action sequence is then fed into the low-level controller, which operates  
72 at 50 Hz.

## 73 **2.5 Evaluation**

74 Here we list the evaluation criteria for each task:

- 75 • Hang Mug (Sim): If the final height of the mug is larger than 0.1m and lower than 0.3m,  
76 and the xy-space distance between the mug holder and the mug is lower than 0.1m, the  
77 reward is 1. Otherwise, the reward is 0.
- 78 • Pencil Insertion (Sim): If the pencil’s final height is larger than 0.05m, and the xy-space an-  
79 gle between the pencil and the hole is smaller than  $\arccos(0.8)$ , the reward is 1. Otherwise,  
80 the reward is 0.
- 81 • Place Can: The policy rollout is only successful if the can is grasped and placed upright on  
82 the table.
- 83 • Use Toothbrush: The policy rollout is only successful if the toothbrush is grasped and the  
84 head of the toothbrush is aligned with the toothpaste tip.
- 85 • Align Shoes: The policy rollout is only successful if the shoe is rotated towards the left.
- 86 • Use Spoon: The policy rollout is only successful if the robot grasps the spoon, scoops the  
87 coffee beans, and pours into the bowl.
- 88 • Collect Knife: The policy rollout is only successful if the robot grasps the knife handle and  
89 puts it in the box.
- 90 • Open Pen: The policy rollout is only successful if the robot first grasps the pen body using  
91 one hand, and then pulls out the pen tap using another hand.

# 92 **3 Additional Experiments**

## 93 **3.1 Quantitative Experiment Table**

94 As mentioned in Section 4.2 of the main paper, we present a detailed table for Figure 4. This table  
95 provides more concrete results regarding our comparisons with the baseline methods. We can see

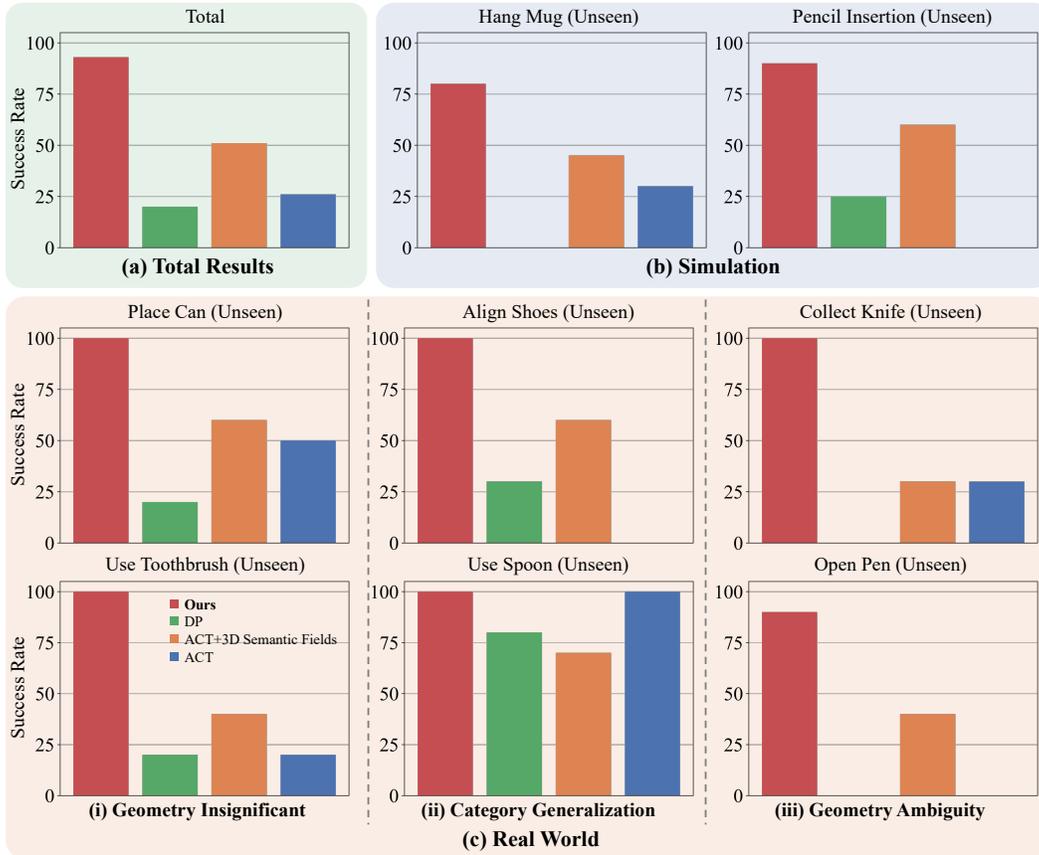


Figure 1: **Success Rate of Extension to ACT.** We measure the success rate of our method, diffusion policy, ACT with our 3D semantic fields, and ACT on unseen instances across different tasks. We found that our 3D semantic fields can help other imitation learning methods like ACT to generalize to novel instances as well. However, in practice, we choose the diffusion policy with 3D semantic fields since it performs the best across different tasks.

96 that, although the diffusion policy performs well for the seen instances, it fails to generalize to  
 97 novel instances. Compared with our method without semantics, our method also demonstrates the  
 98 effectiveness of incorporating semantic information.

### 99 3.2 Extension to ACT

100 We also extend our representation to other imitation learning methods, such as ACT [3]. Figure 1  
 101 shows the experiment results. First, we found that our representation helps other imitation learning  
 102 algorithms generalize to novel instances. This is because raw RGB images are sensitive to environ-  
 103 mental factors and object appearances, while our 3D semantic fields are robust against environmental  
 104 changes and novel instances by explicitly using geometric and semantic information. Second, we  
 105 found that our method outperforms ACT. We believe the main reason is that the diffusion policy is  
 106 better at capturing multi-modal demonstration distribution compared to ACT, which is quite com-  
 107 mon in the real world. In summary, our representation can be applied to different imitation learning  
 108 methods, but we choose the diffusion policy because it performs the best in practice.

### 109 3.3 Ablation on Feature Extraction Backbone

110 Originally, we used DINOv2 as the feature backbone [4]. To study the influence of different feature  
 111 backbones on performance, we replace DINOv2 with other feature backbones. Other than DINOv2,  
 112 CLIP and Vision Transformer (ViT) are two of the most common image feature extractors [5, 6].  
 113 Therefore, we replace DINOv2 with CLIP and ViT and compare their performance. First, we try to

Task Category Task Name Instances	Simulation			
	Hang Mug		Insert Pencil	
	Seen	Unseen	Seen	Unseen
Ours	<b>95% (19/20)</b>	80% (16/20)	<b>95% (19/20)</b>	<b>90% (18/20)</b>
Ours w/o Semantics	65% (13/20)	<b>85% (17/20)</b>	40% (8/20)	40% (8/20)
Diffusion Policy	<b>95% (19/20)</b>	0% (0/20)	<b>95% (19/20)</b>	25% (5/20)
Diffusion Policy w/ RGBD	0% (0/20)	0% (0/20)	15% (3/20)	20% (4/20)

Task Category Task Name Instances	Geometry Ambiguity		Geometry Details	
	Collect Knife	Open Pen	Place Can	Use Toothbrush
	Unseen	Unseen	Unseen	Unseen
Ours	<b>100% (9/10)</b>	<b>90% (9/10)</b>	<b>100% (10/10)</b>	<b>100% (10/10)</b>
Ours w/o Semantics	50% (5/10)	50% (5/10)	60% (6/10)	30% (3/10)
Diffusion Policy	0% (0/10)	0% (0/10)	20% (2/10)	20% (2/10)
Diffusion Policy w/ RGBD	60% (6/10)	20% (2/10)	0% (0/10)	20% (2/10)

Task Category Task Name Instances	Category Generalization		Total (Seen)	Total (Unseen)
	Collect Knife	Open Pen		
	Unseen	Unseen		
Ours	<b>100% (10/10)</b>	<b>100% (10/10)</b>	<b>95% (38/40)</b>	<b>93% (93/100)</b>
Ours w/o Semantics	60% (6/10)	90% (9/10)	53% (21/40)	59% (59/100)
Diffusion Policy	30% (3/10)	80% (8/10)	<b>95% (38/40)</b>	20% (20/100)
Diffusion Policy w/ RGBD	10% (1/10)	70% (7/10)	8% (3/40)	22% (22/100)

Table 2: **Success Rate.** The method was evaluated across eight tasks. We observe that the diffusion policy performs similarly to our method in the seen environments but shows markedly worse performance in unseen instances. In addition, compared with ours without semantics, our method shows a better performance across different tasks and demonstrates the necessity of semantic information. These results underscore our policy’s capability to achieve category-level generalization and encode semantic information.



Figure 2: **3D Semantic Fields Visualization.** We visualize 3D semantic fields in the scene, similar to Figure 6 in the main paper. For ours with DINOv2, object parts from different instances are highlighted consistently. For example, shoe heads, book titles, and mug handles are highlighted in these scenes. In contrast, ours with CLIP and ours with ViT fail to highlight object parts consistently. Instead, their heatmap distributes over the scene and has no distinctions among object parts.

114 qualitatively understand what the 3D semantic fields will look like with different feature backbones.  
 115 Second, we quantitatively show the performance of the whole diffusion policy pipeline when the  
 116 feature backbones are different.

117 Similar to Figure 6 in the main paper, we use the same 3D semantic field construction and visual-  
 118 ization pipeline with different feature backbones. We visualize the 3D semantic fields of different  
 119 object parts in Figure 2. We found that ours with DINOv2 can consistently highlight the object parts,  
 120 such as shoe heads, book titles, and mug handles, while ours with CLIP and ours with ViT fail to do  
 121 so. Without the capability to distinguish semantically meaningful parts of the object, ours with ViT  
 122 and ours with CLIP fail to generalize to unseen instances, distinguish geometric ambiguities, and  
 123 attend to subtle geometric details. This qualitative result justifies our usage of DINOv2 to generate  
 124 useful 3D semantic fields.

125 Table 3 shows our quantitative results, where the experiment process is the same as the main paper.  
 126 The table shows our method’s average success rate is 94%, which is the best among all methods. As  
 127 this table shows, ours with CLIP’s average success rate is 57%, and ours with ViT’s average success  
 128 rate is 55%. This result is close to ours without semantics (58%) as shown in Table 1. This indicates  
 129 that CLIP and ViT do not encode useful semantic information like object parts. In contrast, our  
 130 method can highlight parts of the object that are critical for the task’s success.

Task Category	Task Name	Instances	Ours	Ours with CLIP	Ours with ViT
Sim.	Hang Mug	Seen	<b>95% (19/20)</b>	55% (11/20)	80% (16/20)
		Unseen	<b>80% (16/20)</b>	50% (10/20)	70% (14/20)
	Insert Pencil	Seen	<b>95% (19/20)</b>	70% (14/20)	55% (11/20)
		Unseen	<b>90% (18/20)</b>	45% (9/20)	30% (6/20)
Geo. Ambi.	Collect Knife	Unseen	<b>100% (10/10)</b>	30% (3/10)	30% (3/10)
	Open Pen	Unseen	<b>100% (10/10)</b>	50% (5/10)	40% (4/10)
Geo. Details	Place Can	Unseen	<b>100% (10/10)</b>	40% (4/10)	50% (5/10)
	Use Toothbrush	Unseen	<b>90% (9/10)</b>	50% (5/10)	50% (5/10)
Cat. Gen.	Align Shoes	Unseen	<b>100% (10/10)</b>	90% (9/10)	40% (4/10)
	Use Spoon	Unseen	<b>100% (10/10)</b>	100% (10/10)	90% (9/10)
Total (Unseen)			<b>94% (131/140)</b>	57% (80/140)	55% (77/140)

Table 3: **Success Rate for Different Feature Backbones.** The method was evaluated across eight tasks and compared against different feature backbones. We use the average success rate as our evaluation metric. Compared with ours with CLIP and ours with ViT, our method consistently outperforms them. This demonstrates that DINOv2 could encode semantic information, such as parts of object that are important for task completion.

### 131 3.4 More Baselines

132 We compare with SORNet and DINOBot [7, 8]. We will describe how we compare our work with  
 133 these two baselines in our setting respectively in Section 3.4.1 and Section 3.4.2. Then we present  
 134 results and analyze them in Section 3.4.3.

#### 135 3.4.1 SORNet

136 SORNet introduces an object-centric representation that can be generalized to objects unseen during  
 137 training. Specifically, SORNet can take in observation from more than one view, where depth is  
 138 optional. It also takes in canonical object views and outputs corresponding embedding for each  
 139 query object. For fair comparisons, we input RGBD observations from four viewpoints and one  
 140 canonical object view into SORNet and obtain one embedding. Then we input this embedding to  
 141 diffusion policy, the same as what did for our approach. We do not use pre-trained models provided

142 by SORNet, since the observation modality and training domain are very different. Instead, we train  
 143 the whole model from scratch.

### 144 3.4.2 DINOBot

145 DINOBot introduces a single-shot imitation learning framework that leverages visual features from  
 146 foundational vision models. They first record the key end-effector pose and corresponding RGBD  
 147 observations. Then they record a motion sequence. During testing, they will align current RGBD  
 148 observations with recorded key RGBD observations to make sure the end-effector has arrived at the  
 149 key pose. Then they will replay the motion sequence.

150 In our implementation, we choose one of our perspective views as the observation view. For the rest  
 151 of the implementation, we follow the DINOBot.

### 152 3.4.3 Results

153 As shown in Table 4, SORNet performs badly for both seen and unseen instances. As mentioned  
 154 by SORNet’s authors, SORNet needs to be trained on a large-scale dataset with diverse object ap-  
 155 pearance and scene layouts. Specifically, they trained SORNet on two datasets, which contain 405  
 156 training objects and 330 training objects respectively. However, we only assume access to a demon-  
 157 stration dataset collected over one or a handful of objects. SORNet cannot effectively extract useful  
 158 representation in this low-data regime, while our representation could generalize to unseen instances  
 159 given the same amount of data.

Task Category	Task Name	Instances	Ours	SORNet [7]	DINOBot [8]
Sim.	Hang Mug	Seen	<b>95% (19/20)</b>	0% (0/20)	0% (0/20)
		Unseen	<b>80% (16/20)</b>	0% (0/20)	0% (0/20)
	Insert Pencil	Seen	<b>95% (19/20)</b>	0% (0/20)	0% (0/20)
		Unseen	<b>90% (18/20)</b>	0% (0/20)	0% (0/20)
Geo. Ambi.	Collect Knife	Unseen	<b>100% (10/10)</b>	10% (1/10)	0% (0/10)
	Open Pen	Unseen	<b>100% (10/10)</b>	0% (0/10)	N/A% (N/A)
Geo. Details	Place Can	Unseen	<b>100% (10/10)</b>	20% (2/10)	0% (0/10)
	Use Toothbrush	Unseen	<b>90% (9/10)</b>	0% (0/10)	0% (0/10)
Cat. Gen.	Align Shoes	Unseen	<b>100% (10/10)</b>	0% (0/10)	0% (0/10)
	Use Spoon	Unseen	<b>100% (10/10)</b>	40% (4/10)	0% (0/10)
Total (Unseen)			<b>94% (131/140)</b>	5% (7/140)	0% (0/130)

Table 4: **Success Rate for Additional Representation and Imitation Learning Baselines.** The method was evaluated across eight tasks and compared with SORNet and DINOBot. We use the average success rate as our evaluation metric. The low success rate of SORNet implies that it needs large-scale data for pre-training, while we only assume access to demonstrations recorded in several hours. DINOBot’s low success rate implies it is not as flexible as our method for action representation and task specification. There is no result for the opening pen task because of DINOBot’s assumption of single-arm manipulation.

160 As shown in Table 4, DINOBot cannot accomplish the tasks and fails in both seen and unseen  
 161 scenarios. The reasons are twofold. First, DINOBot merely replays the recorded trajectory after the  
 162 6-DoF alignment, which limits its applicability in tasks that require adaptive adjustments based on  
 163 the positions and shapes of objects. Second, DINOBot has only tackled single-arm manipulation  
 164 tasks using a wrist-mounted camera, which restricts its extension to more complicated bimanual  
 165 tasks (e.g., pen opening, as we demonstrated using our method). In contrast, our method represents  
 166 the action as a sequence of end-effector poses, which can be easily extended to a dual-arm setup and  
 167 accomplish a more diversified set of tasks.

168 **3.5 t-SNE**

169 In addition, we also analyze whether our se-  
 170 mantic fields could carry desired semantic in-  
 171 formation, as shown in Figure 3. We ini-  
 172 tially sample 3D grid points in the workspace  
 173 and select the top 100 points with the highest  
 174 similarity score for each semantic field. For  
 175 the example shown in Figure 3, there are 500  
 176 points in total. Then, we query the descrip-  
 177 tor fields and obtain semantic features corre-  
 178 sponding to the selected points. We reduce the  
 179 high-dimensional semantic features to a two-  
 180 dimensional plane using t-distributed stochas-  
 181 tic neighbor embedding (t-SNE). We observe  
 182 that all points are distinctly separated, and each  
 183 blob corresponds to one object part. This vi-  
 184 sualization shows that our semantic fields have  
 185 a clear semantic meaning for each channel,  
 186 which helps the policy to achieve category-  
 187 level generalization.

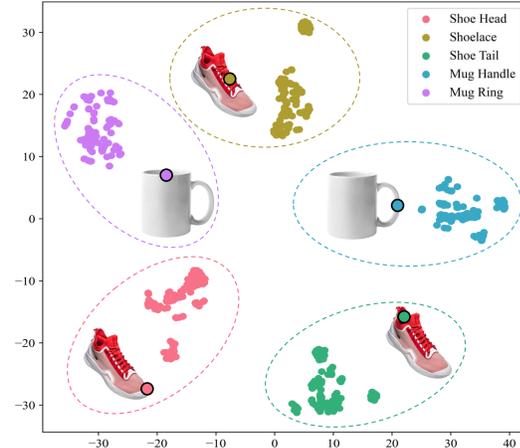


Figure 3: **t-SNE for Semantic Features with High Similarity.** Points in similarity fields are selected based on the top  $k$  similarity scores. We compute their corresponding semantic features and project them into two-dimensional space using t-SNE. It is clearly observed that all features are clustered and separated from each other. Each cluster corresponds to a feature that has semantic meaning, such as mug rings and mug handles.

188 **4 Detailed Comparisons with Prior Works**

- 189 • *Functional Object-Oriented Network for Manipulation Learning* [9]: This work proposes  
 190 a structured knowledge representation and generates motion sequences based on the repre-  
 191 sentation. **While this work represents the motion sequence as motion harmonics, our**  
 192 **work is more flexible in action and task representation. In addition, this work does not**  
 193 **show extensive real-world deployment results, while we focus more on the real-world**  
 194 **deployment.**
- 195 • *Affordance Detection for Task-Specific Grasping Using Deep Learning* [10]: This work  
 196 uses convolutional neural networks to detect object affordance, class, and orientation for  
 197 task-specific tasks. **Although it shows generalization capabilities, it is still restricted to**  
 198 **grasping tasks. In contrast, our framework supports a wider range of tasks.**
- 199 • *SORNet: Spatial Object-Centric Representations for Sequential Manipulation* [7]: SOR-  
 200 Net proposes an object-centric representation from RGB observations. We note three ma-  
 201 jor differences from SORNet to our work. **First, SORNet is object-centric, while ours**  
 202 **is scene-representation, which needs less prior information. Second, SORNet builds**  
 203 **representation from RGB, which makes it hard to generalize to novel domains, envi-**  
 204 **ronments, and backgrounds, but our 3D semantic fields show category-level general-**  
 205 **ization capabilities.** Lastly, through experiments, we found that SORNet might be better  
 206 if pre-trained on a large dataset. **In our work, we only assume access to a demonstra-**  
 207 **tion dataset of one or a handful of objects, while SORNet needs a larger dataset for**  
 208 **pre-training to have a good performance.**
- 209 • *Manipulation-Oriented Object Perception in Clutter Through Affordance Coordinate*  
 210 *Frames* [11]: This work introduces the Affordance Coordinate Frame (ACF) for category-  
 211 level pose estimation. It shows impressive category-level generalization capabilities. **How-**  
 212 **ever, this work uses pre-defined motion primitives (grasp, pour, stir) for manipulation,**  
 213 **which limits applicable tasks. In contrast, we do not assume any prior motion primi-**  
 214 **tives, which is more flexible for action representation.**
- 215 • *StructDiffusion: Object-Centric Diffusion for Semantic Rearrangement of Novel Ob-*  
 216 *jects* [12]: This work uses a diffusion model and object-centric transformer to construct  
 217 goal structures given current point clouds and language inputs. They mainly focus on pick-

ing and placing tasks using modular methods, including the grasping planner and motion planner. **Their modular method might not transfer to other tasks and other types of objects like deformable objects and granular objects easily. Without assumptions for target tasks and objects, our framework is flexible for object types and task specifications.**

- *A Survey of Semantic Reasoning Frameworks for Robotic Systems* [13]: We position our work as instance-level semantic reasoning according to this survey. In our work, we propose 3D semantic fields, which can differentiate the object parts using foundational vision models.
- *Learning Generalizable Feature Fields for Mobile Manipulation* [14]: GeFF is a concurrent work using Generalizable NeRF for mobile manipulation [15]. There are two major differences between our work and GeFF. **First, we do not distill feature fields from foundational vision models to avoid the loss in generalization capabilities. Second, GeFF uses open-push-close action sequences to accomplish tasks, while our policy predicts the end-effector trajectory, which is more flexible.**
- *Evaluating Robustness of Visual Representations for Object Assembly Task Requiring Spatio-Geometrical Reasoning* [16]: This work mainly focuses on object assembly tasks by evaluating the robustness of visual representations from RGB observations. **First, since RGB observation is sensitive to environment and instance variance, our method proposes to use 3D semantic fields to help the generalization of diffusion policy. Second, instead of focusing on assembly tasks, our framework is tested on a broader set of tasks.**
- *You Only Demonstrate Once: Category-Level Manipulation from Single Visual Demonstration* [17]: You Only Demonstrate Once introduces the category-level generalizable last-inch policy using model-free pose estimation and trajectory transformation. There are two major differences. **First, their behavior cloning algorithm is replaying the trajectory using the object’s pose, which limits their applicability to tasks requiring adaptive adjustments based on the positions and shapes of objects. In contrast, we represent the trajectory as a sequence of end-effector poses, which is more flexible for task specification. In addition, they assume the object is rigidly attached to the end-effector, which is not applicable to some tasks like pushing the shoe.** In summary, many of our tasks are not doable using this framework.
- *On the Effectiveness of Retrieval, Alignment, and Replay in Manipulation and DINOBot: Robot Manipulation via Retrieval and Alignment with Vision Foundation Models* [8, 18]: This line of work from the set of authors proposes a generalizable imitation learning framework, following phases of retrieval, alignment, and replay. DINOBot is the latest work enhancing the prior work using vision foundation models. **First, DINOBot merely replays the recorded trajectory after the 6-DoF alignment, which limits its applicability in tasks that require adaptive adjustments based on the positions and shapes of objects. Second, DINOBot has only tackled single-arm manipulation tasks using a wrist-mounted camera, which restricts its extension to more complicated bimanual tasks (e.g., pen opening, as we demonstrated using our method). In contrast, our method represents the action as a sequence of end-effector poses, which can be easily extended to a dual-arm setup and accomplish a more diversified set of tasks.**

## References

- [1] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.
- [2] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

- 268 [3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation  
269 with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- 270 [4] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez,  
271 D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li,  
272 W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal,  
273 P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without  
274 supervision, 2023.
- 275 [5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell,  
276 P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from  
277 natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th In-*  
278 *ternational Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learn-*  
279 *ing Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL [https://proceedings.](https://proceedings.mlr.press/v139/radford21a.html)  
280 [mlr.press/v139/radford21a.html](https://proceedings.mlr.press/v139/radford21a.html).
- 281 [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. De-  
282 hghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is  
283 worth 16x16 words: Transformers for image recognition at scale. In *International Confer-*  
284 *ence on Learning Representations*, 2021. URL [https://openreview.net/forum?](https://openreview.net/forum?id=YicbFdNTTy)  
285 [id=YicbFdNTTy](https://openreview.net/forum?id=YicbFdNTTy).
- 286 [7] W. Yuan, C. Paxton, K. Desingh, and D. Fox. Sornet: Spatial object-centric representations  
287 for sequential manipulation. In *5th Annual Conference on Robot Learning*, pages 148–157.  
288 PMLR, 2021.
- 289 [8] N. D. Palo and E. Johns. Dinobot: Robot manipulation via retrieval and alignment with vision  
290 foundation models. In *IEEE International Conference on Robotics and Automation (ICRA)*,  
291 2024.
- 292 [9] D. Paulius, Y. Huang, R. Milton, W. D. Buchanan, J. Sam, and Y. Sun. Functional object-  
293 oriented network for manipulation learning. In *2016 IEEE/RSJ International Conference on*  
294 *Intelligent Robots and Systems (IROS)*, pages 2655–2662. IEEE, 2016.
- 295 [10] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic. Affordance detection for task-specific  
296 grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid*  
297 *Robotics (Humanoids)*, pages 91–98. IEEE, 2017.
- 298 [11] X. Chen, K. Zheng, Z. Zeng, C. Kisailus, S. Basu, J. Cooney, J. Pavlasek, and O. C. Jenkins.  
299 Manipulation-oriented object perception in clutter through affordance coordinate frames. In  
300 *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 186–  
301 193. IEEE, 2022.
- 302 [12] W. Liu, Y. Du, T. Hermans, S. Chernova, and C. Paxton. Structdiffusion: Language-guided  
303 creation of physically-valid structures using unseen objects. In *RSS 2023*, 2023.
- 304 [13] W. Liu, A. Daruna, M. Patel, K. Ramachandruni, and S. Chernova. A survey of semantic  
305 reasoning frameworks for robotic systems. *Robotics and Autonomous Systems*, 159:104294,  
306 2023.
- 307 [14] R.-Z. Qiu, Y. Hu, G. Yang, Y. Song, Y. Fu, J. Ye, J. Mu, R. Yang, N. Atanasov, S. Scherer,  
308 and X. Wang. Learning generalizable feature fields for mobile manipulation. *arXiv preprint*  
309 *arXiv:2403.07563*, 2024.
- 310 [15] J. Ye, N. Wang, and X. Wang. Featurenerf: Learning generalizable nerfs by distilling founda-  
311 tion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,  
312 pages 8962–8973, 2023.

- 313 [16] C. Ku, C. Winge, R. Diaz, W. Yuan, and K. Desingh. Evaluating robustness of visual rep-  
314 resentations for object assembly task requiring spatio-geometrical reasoning. *arXiv preprint*  
315 *arXiv:2310.09943*, 2023.
- 316 [17] B. Wen, W. Lian, K. Bekris, and S. Schaal. You only demonstrate once: Category-level ma-  
317 nipulation from single visual demonstration. *Robotics: Science and Systems 2022*, 2022.
- 318 [18] N. Di Palo and E. Johns. On the effectiveness of retrieval, alignment, and replay in manipula-  
319 tion. *IEEE Robotics and Automation Letters*, 2024.