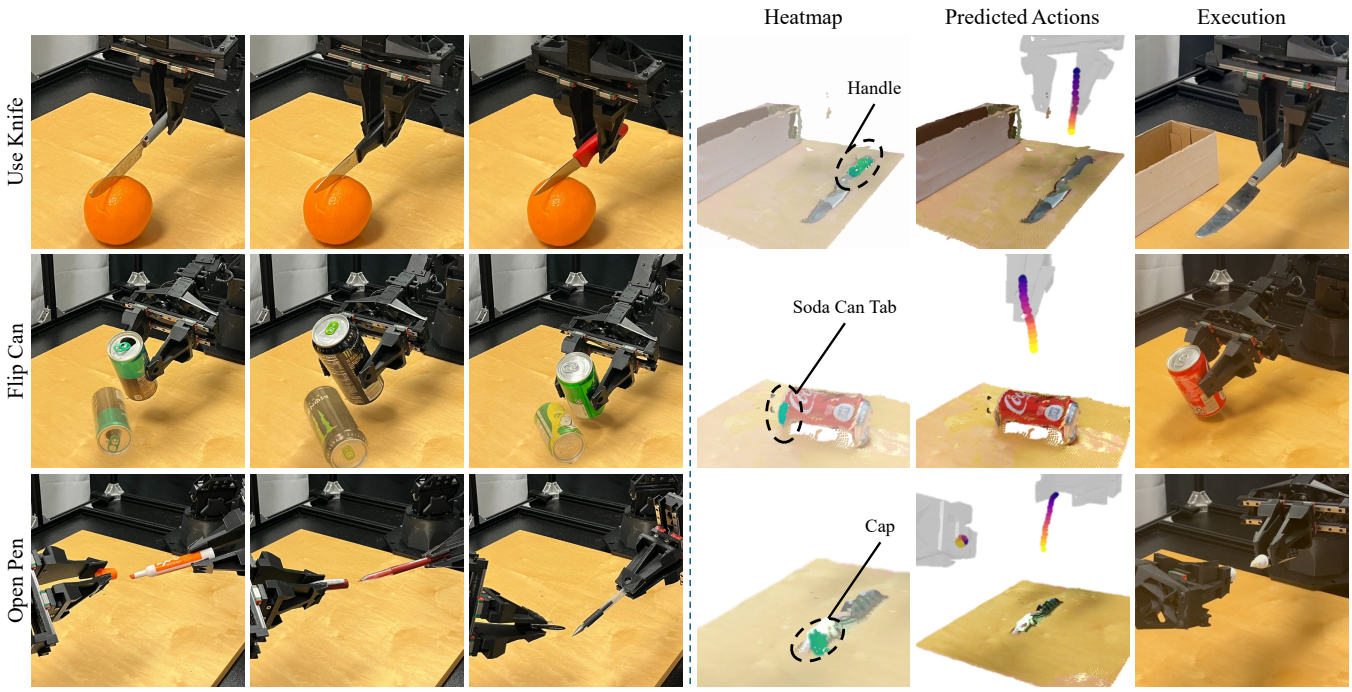
GILD: Generalizable Imitation Learning with 3D Semantic Fields

Yixuan Wang[1], Binghao Huang[1], Guang Yin[1], Tarik Kelestemur[2], Yunzhu Li[1]



(a) Daily Tasks Examples (b) Our Imitation Learning Framework

Fig. 1: **Generalizable Imitation Learning using 3D Semantic Fields.** (a) In daily life, the semantics of object parts are important for task completion. For instance, when using a knife, the robot needs to identify the knife handle and grasp it. When flipping a soda can without spilling out liquids, the robot must locate the soda can tab to put it upright. If not, the soda can might be upside down. Semantic understanding of the environment is essential for task completion and helps to generalize to novel instances. (b) We use 3D semantic fields to build our generalizable imitation learning framework. Our 3D semantic fields could highlight semantically meaningful parts, as shown in the heatmap example. The two columns on the right show our policy's predicted actions and their actual execution results, demonstrating our policy's capability to attend to the right semantic features and accomplish tasks.

***Abstract*— Imitation learning has shown remarkable capability in executing complex robotic manipulation tasks. However, existing frameworks often fall short in structured modeling of the environment, lacking explicit characterization of geometry and semantics, which limits their ability to generalize to unseen objects and layouts. To enhance the generalization capabilities of imitation learning agents, we introduce a novel framework in this work, incorporating explicit spatial and semantic information via 3D semantic fields. We begin by generating 3D descriptor fields from multi-view RGBD observations with the help of large foundational vision models. These high-dimensional descriptor fields are then converted into low-dimensional semantic fields, which aids in the efficient training of a diffusion-based imitation learning policy. The proposed method offers explicit consideration of geometry and semantics, enabling strong generalization capabilities in tasks that require category-level generalization, resolving geometric ambiguities, and attention to subtle geometric details. We evaluate our method across eight tasks involving articulated objects and instances with varying shapes and textures from multiple object categories. Our method proves its effectiveness by outperforming state-of-the-art imitation learning baselines on unseen testing instances by 57%. Additionally, we provide a detailed analysis and visualization to interpret the sources of performance gain and explain how our method can generalize to novel instances.**

[1]University of Illinois Urbana-Champaign [2]Boston Dynamics AI Institute

I. Introduction

Imitation learning has recently shown promising results in real robot deployment for complex robotic manipulation tasks [4, 49]. However, most of the existing end-to-end imitation learning frameworks are brittle to environment variances, such as novel object instances, camera viewpoints, and background changes. When the aforementioned factors change, they often need to collect new demonstrations and train the policy, which is sample inefficient. Previous efforts to tackle these challenges involve extracting geometric information from high-dimensional RGBD observation [51]. Although they are sample-efficient, their sole reliance on geometric information is insufficient. For instance, as illustrated in Figure 1b, a marker's head and tail are geometrically

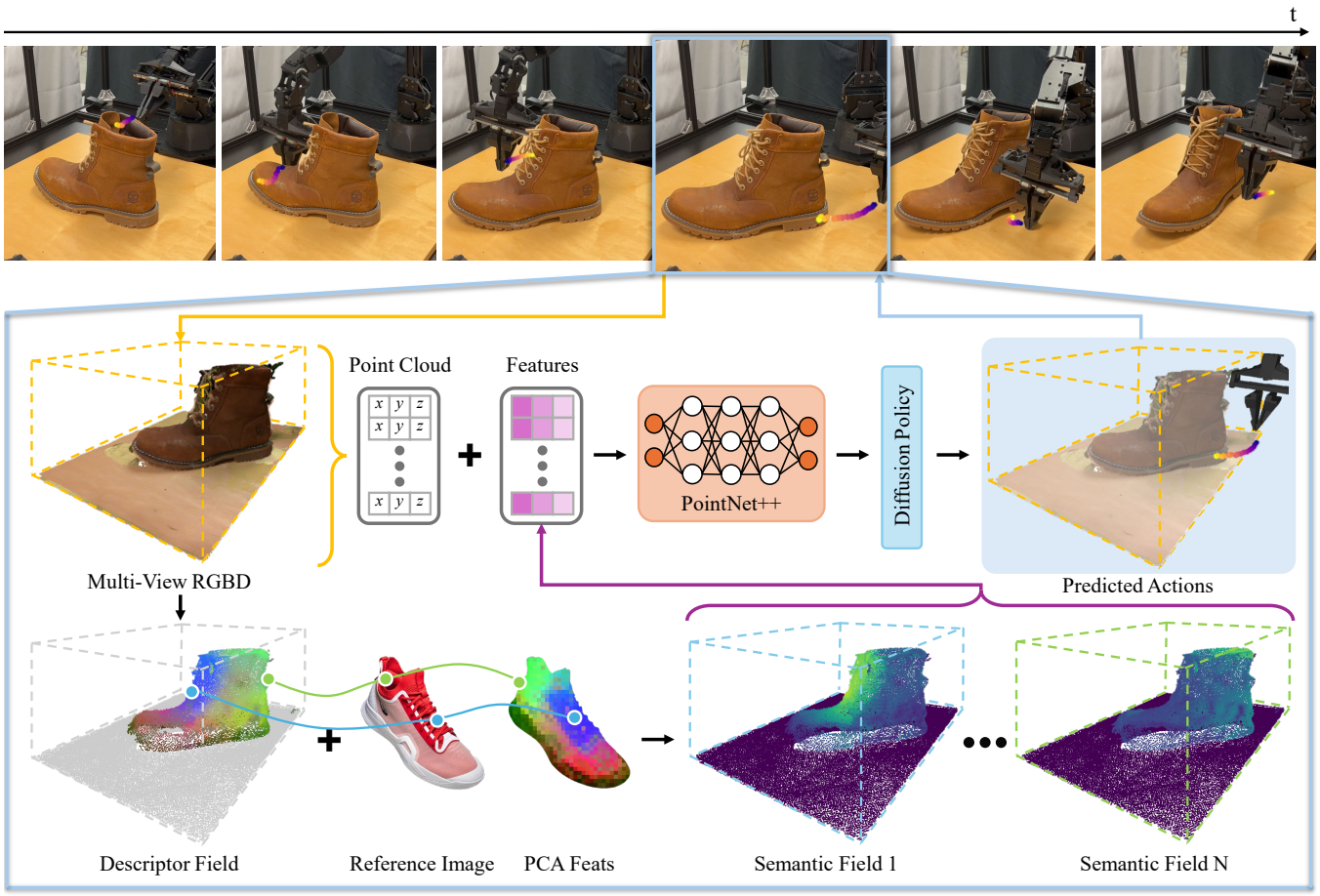


Fig. 2: **Method Overview.** The top row shows a sequence of real policy rollouts in the aligning shoe task. At one time step, we take in multi-view RGBD observations first and then extract the 3D descriptor field, with each point possessing a corresponding high-dimensional descriptor. We then select reference features from 2D reference images. By computing the cosine similarity between the descriptor field and 2D reference semantic features, we could obtain several similarity fields. These similarity fields, concatenated with the point cloud, are then input into PointNet++ and the diffusion policy to output predicted actions.

ambiguous despite their functional and semantic differences.

Therefore, an ideal representation should not only extract geometric information from raw observation ensuring sample efficiency while also retaining semantic information for better and more robust generalization. In this work, we introduce a novel imitation learning framework that uses a scene representation in the form of **3D semantic fields**. Our framework consists of three main modules: a 3D descriptor fields encoder, a semantic fields constructor, and an action policy. The 3D descriptor encoder takes in multi-view RGBD observations. For arbitrary 3D points, it evaluates the associated high-dimensional descriptor using large foundational vision models like DINOv2 [27]. These descriptors are then fed into the semantic fields constructor and converted into low-dimensional semantic fields. Finally, the policy takes in the semantic fields along with the point cloud and predicts actions.

This framework offers three benefits: (1) **Category-level generalization:** As semantic fields contain both 3D and semantic information, it guides our policy to focus on semantically meaningful parts essential for task completion, allowing generalization across instances within a category. (2) **Resolve geometric ambiguity:** Geometric information can be ambiguous. For example, the knife blade and knife handle are geometrically similar despite functional and semantical differences. Our semantic fields can localize space semantically close to parts that are important for task completion, such as knife blades, to disambiguate vague geometric information. (3) **Attention to subtle geometric information**: Geometric details might be lost due to real-world observation noise, such as toothbrush head and soda can tab. Without sufficient details, some tasks are impossible to accomplish, such as spreading toothpaste on a toothbrush and flipping a lying soda can upwards, as shown in Figure 1. Because our semantic fields highlight semantically distinct regions, our method can pay attention to nuanced geometric details for task completion.

We systematically evaluate our method across eight tasks. Our task settings involve novel instances, ambiguous object geometry, and inconspicuous geometric information. The results demonstrate that our approach not only generalizes effectively to new instances but also clarifies geometric ambiguities and enhances subtle geometric information. Compared with baseline methods, which frequently fail to

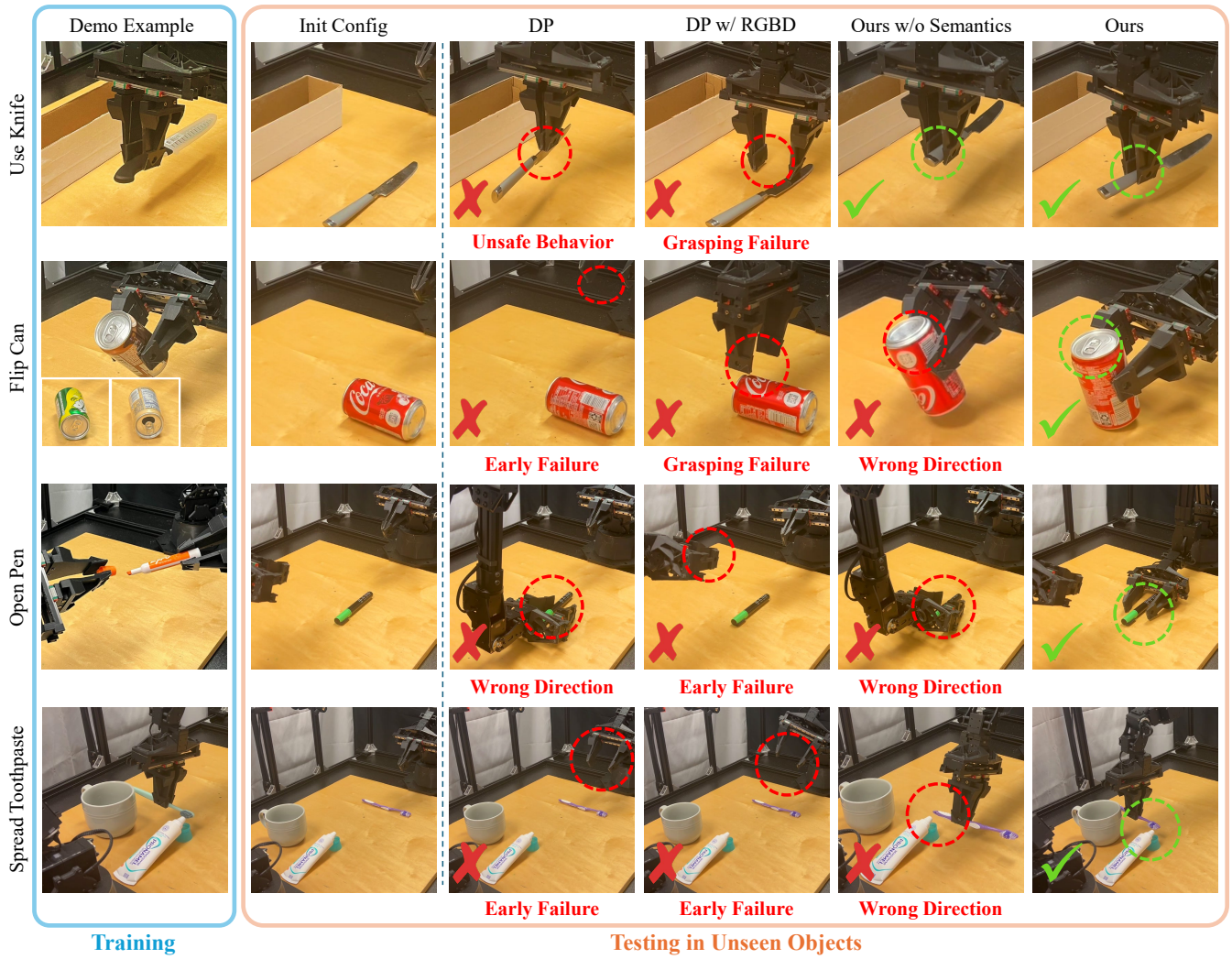


Fig. 3: **Policy Rollout in Real World.** The figure illustrates the policy rollout results in the real-world scenario. On the left, the blue block displays the demonstration examples and corresponding training instances. On the right, the orange block presents the example of policy rollout results. From left to right, they are, respectively, initial configurations, diffusion policy, diffusion policy with RGBD, ours without semantics, and our method. We summarize four common failure modes. Early failure and grasping failure could happen when the novel instance is presented. Diffusion policy may also lead to unsafe behavior when encountering novel instances. Ours without semantics might identify wrong directions due to geometric ambiguity and nuanced geometric details.

generalize to novel instances because they lack both 3D and semantic information in their representation, our approach demonstrates significant advantages. Furthermore, methods based on 3D representations usually fail when the geometric information contains ambiguity or subtle details, which could lead to undesired robot behavior.

In this work, our contributions are threefold: (1) We propose an imitation learning framework that uses semantic fields, which encode the environment's geometric and semantic information. (2) We conduct comprehensive experiments and suggest that our method can generalize to novel instances, resolve geometric ambiguity, and amplify inconspicuous geometric information. Our method surpasses the best baseline method by 57%. (3) We provide a detailed analysis of how well and why our method can generalize to novel instances.

II. Method

This section first presents our problem formulation in Section IV-B.1. We then describe how to extract descriptor fields from raw observations in Section IV-B.2, as illustrated in the bottom left of Figure 2. Next, we explain the construction of semantic fields, as displayed in the bottom right of Figure 2, in Section IV-B.3. Finally, we discuss how to learn the policy to predict actions given our representation in Section IV-B.4.

III. Experiments

In this section, we evaluate our method on eight diverse tasks. We aim to answer the following three questions through experiments. 1) How does the performance of our method compare to that of the state-of-the-art methods? 2) How well can our method generalize to different configurations and different instances? 3) What enabled our method to generalize to novel instances?

References

[1] Dominik Bauer, Timothy Patten, and Markus Vincze. Reagent: Point cloud registration using imitation and reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14586–14594, 2021.

[2] Haonan Chen, Yilong Niu, Kaiwen Hong, Shuijing Liu, Yixuan Wang, Yunzhu Li, and Katherine Rose Driggs-Campbell. Predicting object interactions with behavior primitives: An application in stowing tasks. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=VH6WIPF4Sj.

[3] Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 8 (84):eadc9244, 2023. doi: 10.1126/scirobotics.adc9244. URL https://www.science.org/doi/abs/10.1126/scirobotics.adc9244.

[4] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[5] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

[6] Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. *Advances in neural information processing systems*, 30, 2017.

[7] Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. *Conference on Robot Learning (CoRL)*, 2021.

[8] Peter Florence, Lucas Manuelli, and Russ Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5(2): 492–499, 2019.

[9] Aditya Ganapathi, Pete Florence, Jake Varley, Kaylee Burns, Ken Goldberg, and Andy Zeng. Implicit kinematic policies: Unifying joint and cartesian action spaces in end-to-end robot learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2656–2662. IEEE, 2022.

[10] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *Conference on Robot Learning*, pages 3949–3965. PMLR, 2023.

[11] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. *arXiv preprint arXiv:2306.14896*, 2023.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[13] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.

[14] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[15] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *Arxiv*, 2024.

[16] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.

[17] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *ICLR*, 2019.

[18] Yen-Chen Lin, Pete Florence, Andy Zeng, Jonathan T Barron, Yilun Du, Wei-Chiu Ma, Anthony Simeonov, Alberto Rodriguez Garcia, and Phillip Isola. Mira: Mental imagery for robotic affordances. In *Conference on Robot Learning*, pages 1916–1927. PMLR, 2023.

[19] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.

[20] Lucas Manuelli, Yunzhu Li, Pete Florence, and Russ Tedrake. Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning. *arXiv preprint arXiv:2009.05085*, 2020.

[21] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

[23] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Igor Mordatch. Concept learning with energy-based models. *arXiv preprint arXiv:1811.02486*, 2018.

[25] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910, 2019.

[26] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. https://octo-models.github.io, 2023.

[27] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[28] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[29] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

[30] Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Hao Su, and Xiaolong Wang. Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In *Conference on Robot Learning*, pages 594–605. PMLR, 2023.

[31] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. *arXiv preprint arXiv:2306.10007*, 2023.

[32] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=k-Fg8JDQmc.

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

[34] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

[35] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

[36] Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. Robocraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks. *arXiv preprint arXiv:2205.02909*, 2022.

[37] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023.

[38] Lucy Xiaoyang Shi, Archit Sharma, Tony Z Zhao, and Chelsea Finn. Waypoint-based imitation learning for robotic manipulation. *arXiv preprint arXiv:2307.14326*, 2023.

[39] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[40] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[41] 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.

[42] Yixuan Wang, Yunzhu Li, Katherine Driggs-Campbell, Li Fei-Fei, and Jiajun Wu. Dynamic-Resolution Model Learning for Object Pile Manipulation. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.047.

[43] Yixuan Wang, Zhuoran Li, Mingtong Zhang, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D^3fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. *arXiv preprint arXiv:2309.16118*, 2023.

[44] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 2011. URL https://api.semanticscholar.org/CorpusID:2178983.

[45] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[46] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang,

Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. *arXiv preprint arXiv:2312.01853*, 2023.

[47] Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on Robot Learning*, pages 284–301. PMLR, 2023.

[48] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635. IEEE, 2018.

[49] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[50] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[51] Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *7th Annual Conference on Robot Learning*, 2023.

IV. Supplementary

A. Related Works

1) Imitation Learning for Robotic Manipulation: There are mainly three types of imitation learning policies for robotic manipulation. The first one is the explicit policy, which takes observations as inputs and outputs actions and outputs actions directly [6, 8, 11, 19, 33, 38–40, 48–50]. The supervision signal is usually from the demonstration actions. However, this approach often struggles with modeling the multi-modal distribution of demonstrations, which are commonly seen in the real world.

To tackle the challenge of modeling multi-modal demonstrations, another approach models the policy as an implicit function [5, 7, 9, 24, 34, 44]. Instead of predicting actions directly, they build the Energy-Based Models (EBMs) and assign actions with energy values. During testing, an optimizer is used to find the optimal actions. This approach enables the modeling of multi-modal distributions of the demonstration actions due to the nature of the energy model.

Prior diffusion-based policy learning studies argue that implicit policy training can be unstable [4, 26, 29]. Unlike implicit policy, diffusion-based policy predicts the noise given current observations and noisy actions. Prior works demonstrate their ability to handle multi-modal distributions of demonstrations and accomplish complicated tasks in the real world. Thus, we model our policy as a diffusion model conditioned on current observation.

However, the majority of aforementioned works rely on 2D observations, which are not robust to environmental changes, such as lighting conditions, camera viewpoints, background changes, and new instances. In contrast to previous works, our method leverages explicit spatial and semantic information with the help of a large foundation model, DINOv2 [27], to generalize over unseen objects.

2) Scene Representation for Vision-Based Manipulation: Scene representation has been one of the essential components of robotics systems. There exist several types of typical scene representation for robotic manipulation.

One common scene representation will be RGB information, on which a majority of previous work in imitation learning has solely relied [4, 14, 26, 28, 29, 31, 34, 49]. RGB information includes all information from environments, which is sufficient for downstream decision-making. However, this information is vulnerable to environmental changes. By contrast, our framework encodes 3D and semantic information, enabling our policy to better generalize to novel instances.

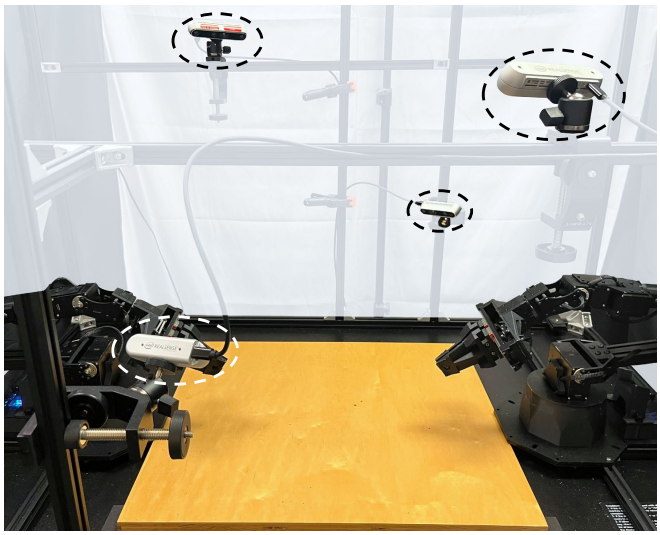
Another research direction represents the scene using point clouds [1, 3, 10, 11, 13, 25, 30, 46, 51]. Using point clouds as the representation enables robots to focus on geometry, facilitating generalization across different environments. Yet, this often results in the loss of critical RGB data, limiting the understanding of an object's semantic properties. Unlike these approaches, our approach retains both semantic and geometric information enhancing downstream decision-making.

Keypoints are another commonly used representation in robotic manipulation [2, 17, 20, 23, 36, 37, 42]. They have shown impressive generalization capabilities. However, they are often sparse and lack rich geometric information, especially for geometric details, limiting the range of tasks that can be effectively addressed. In contrast, our framework employs a representation that encompasses rich geometric information and enables a wide range of tasks.

A recent line of research proposed leveraging neural implicit models such as Occupancy Networks [21] or NeRFs [22] to encode semantic features [15, 16, 18, 25, 32, 35, 41, 43, 47]. Among these, we selected D^3Fields due to its computation efficiency to meet imitation learning's real-time needs. [43] introduces a descriptor field that segments the objects and maps DINOv2 features onto a 3D field, employing model predictive control for action output. However, our methodology differs by embracing an imitation learning policy that greatly benefits from the integration of semantic information.

B. Method

This section first presents our problem formulation in Section IV-B.1. We then describe how to extract descriptor fields from raw observations in Section IV-B.2, as illustrated in the bottom left of Figure 2. Next, we explain the construction of semantic fields, as displayed in the bottom right of Figure 2, in Section IV-B.3. Finally, we discuss how to learn the policy to predict actions given our representation in Section IV-B.4.



(a) Robot Setup

(b) Objects

Fig. 4: **Real Experiment Setup.** (a) We use four RealSense cameras to capture RGBD observations and ALOHA robots to execute policy. (b) We test on a diverse set of objects, including shoes, soda cans, marker pens, knives, spoons, toothbrushes, and toothpaste, with diverse geometry and appearance.

1) Problem Statement: We define our system as a Markov Decision Process (MDP) consisting of state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. The system transition is defined through the dynamics model $s_{t+1} = f(s_t, a_t)$. The goal is to find an optimal policy $a_t = \pi(s_t)$ that can maximize the task reward. Given a set of human demonstrations $D = \{\tau_0, \tau_1, ..., \tau_N\}$, where τ_i represents a trajectory comprising $\{s_0, a_0, s_1, ..., a_T\}$. Here, the state s consists of a sequence of multi-view RGBD observations, while the action a involves a sequence of 3D robot end-effector poses and gripper status.

2) 3D Descriptor Fields: We use an off-the-shelf large foundational vision model, DINOv2, to obtain semantic features from multi-view RGBD observations [27]. Given its ability to extract consistent semantic features from the RGB images across context and instance variances, we selected it as the backbone network.

We provide pseudocode for building 3D descriptor fields in Algorithm 1. We denote single-view RGBD observation as $\boldsymbol{o}_i = (\mathcal{I}_i, \mathcal{R}_i)$, with $i \in \{1, 2, ..., N\}$ representing the camera index, consisting of an RGB image $\mathcal{I}_i \in \mathbb{R}^{H \times W \times 3}$ and a depth image $\mathcal{R}_i \in \mathbb{R}^{H \times W}$. We first use DINOv2 to extract dense 2D feature maps $\mathcal{W}_i$ corresponding to the RGB image $\mathcal{I}_i$ [27]. For an arbitrary 3D point p, we project it onto the image space to find its corresponding pixel location u_i and the distance to camera r_i. We then interpolate to derive features f_i from the feature map and the depth r_i' from $\mathcal{R}_i$.

The depth difference $\Delta r_i = r_i' - r_i$ reflects how distant p is from the surface. When p is closer to the surface in view i, greater weight is given to f_i. We fuse features from multiple viewpoints by applying a weighted sum, thus obtaining the descriptor f corresponding to p. In practice, we follow the implementation details in [43] to extract point cloud $\mathcal{P} \in \mathbb{R}^{K \times 3}$ and associated features $\mathcal{F} \in \mathbb{R}^{K \times F}$, where K is the point cloud's size. We refer readers to [43] for details.

Algorithm 1 3D Descriptor Fields Computation

Infer feature map $\mathcal{W}_i$ from single RGB image $\mathcal{I}_i$
procedure EVALUATE(p)
    Project p to camera i and compute projected pixel u_i and distance to camera r_i
    Obtain interpolated features $f_i = \mathcal{W}_i[u_i]$
    Obtain depth $r_i' = \mathcal{R}_i[u_i]$
    Compute depth difference $\Delta r_i = r_i' - r_i$
    Fuse features $f = h(f_i, \Delta r_i), i \in \{1, 2, ..., N\}$

3) 3D Sematic Fields: A simplistic approach to building representations by concatenating descriptors to raw point clouds has two main issues. Firstly, the storage and computation of high-dimensional descriptors become computationally intensive. For instance, the smallest DINOv2 model variant encodes one image into a feature map with 384 dimensions, resulting in a size increase by more than 100 times compared to the raw point cloud. Secondly, because of the high-dimensional descriptor space, it requires more data to sufficiently cover the descriptor space to reach the desired generalization capabilities.

Instead of directly concatenating descriptors, we use semantic fields to encode semantic information efficiently. First, we select a set of reference descriptors $\mathcal{F}_{\text{ref}} \in \mathbb{R}^{M \times F}$ from 2D images for one object category. The 2D images could be arbitrary images containing the target object category, and each descriptor could represent a part of the object, such as shoe head and shoe tail.

Given the set of reference descriptors, the semantic fields $\mathcal{C} \in \mathbb{R}^{K \times M}$ are defined as the similarity between the descriptor fields and reference descriptors:

$$\mathcal{C}_{ij} = \frac{\mathcal{F}_i \cdot \mathcal{F}_{\text{ref},j}}{||\mathcal{F}_i|| ||\mathcal{F}_{\text{ref},j}||}. \tag{1}$$

By computing the similarity scores, we convert high-dimensional descriptor fields into M-dimensional semantic fields, where M is typically less than 5. We then concatenate semantic fields $\mathcal{C}$ and raw point cloud $\mathcal{P}$ together, which are then inputted into the policy.

Task Category	Task Name	Instances	Ours	Ours w/o Semantics	Diffusion Policy	Diffusion Policy w/ RGBD
Simulation	Hang Mug	Seen	**95% (19/20)**	75% (15/20)	80% (16/20)	0% (0/20)
		Unseen	**85% (17/20)**	75% (15/20)	10% (2/20)	0% (0/20)
	Insert Pencil	Seen	90% (18/20)	40% (8/20)	95% (19/20)	**100% (20/20)**
		Unseen	**80% (16/20)**	35% (7/20)	10% (2/20)	30% (6/20)
Geometry Ambiguity	Collect Knife	Seen	**100% (10/10)**	40% (4/10)	50% (5/10)	0% (0/10)
		Unseen	**100% (10/10)**	80% (8/10)	20% (2/10)	10% (1/10)
	Open Pen	Unseen	**80% (8/10)**	40% (4/10)	10% (1/10)	0% (0/10)
Geometry Insignificant	Place Can	Unseen	**100% (10/10)**	30% (3/10)	10% (1/10)	20% (2/10)
	Use Toothbrush	Unseen	**100% (10/10)**	50% (5/10)	0% (0/10)	0% (0/10)
Category Generalization	Align Shoes	Unseen	**100% (10/10)**	70% (7/10)	30% (3/10)	0% (0/10)
	Use Spoon	Unseen	**100% (10/10)**	90% (9/10)	70% (7/10)	0% (0/10)
Total (Unseen)			**91% (91/100)**	58% (58/100)	23% (23/100)	9% (9/100)

TABLE I: **Success Rate.** The method was evaluated across eight tasks. Our method consistently outperforms the best baseline, except for the pencil insertion task for the seen instances being marginally worse. We observe that ours without semantics performs significantly worse in the presence of ambiguous geometry and subtle geometric details. Furthermore, the diffusion policy performs similarly to our method in the seen environments but shows markedly worse performance in unseen instances. These results underscore our policy's capability to achieve category-level generalization and encode semantic information.

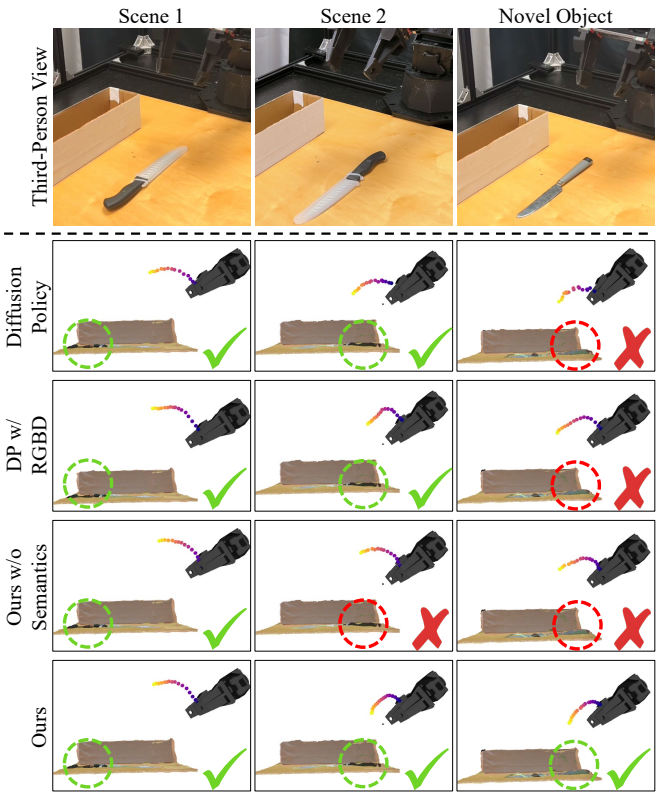


Fig. 5: **Predicted Action Comparisons.** We compare different policies' responses under the same state. For diffusion policy, we observe that its outputs change with knife direction accordingly, but could lead to nonsmooth actions for novel instances. For diffusion policy with RGBD, it might learn to rely on color information more. Therefore, it failed to predict a successful trajectory toward the knife handle. Since geometric information is ambiguous here, our method without semantics fails to distinguish between two knife directions, resulting in similar actions being predicted. When encountering a novel knife, though the predicted action appears reasonable, it still fails to identify the knife handle correctly. In contrast, our method consistently identifies the correct position for the knife handle, regardless of whether the instances are seen or unseen.

4) Policy Learning: In this work, we model our policy as Denoising Diffusion Probabilistic Models (DDPMs), similar to Diffusion Policy [4, 12]. Instead of regressing the action directly, we train a noise predictor network

$$\widehat{\epsilon^k} = \epsilon_\theta(a^k, s, k), \tag{2}$$

that takes in noisy actions a^k, current observations s, and denoising iterations k and predicts the noise $\widehat{\epsilon^k}$. During training, we randomly choose a denoising step k and sample noise ϵ^k added to the unmodified sample a^0. Our training objective is to minimize the difference between ϵ^k and predicted noise:

$$\mathcal{L} = \text{MSELoss}(\epsilon^k, \widehat{\epsilon^k}). \tag{3}$$

During the inference time, our policy starts from random actions a^K and denoises for K steps to obtain the final action predictions. At each step, the action is updated following

$$a^{k-1} = \alpha\big(a^k - \gamma\epsilon_\theta(a^k, s, k) + \mathcal{N}(0, \sigma^2 I)\big), \tag{4}$$

where α, γ, and σ are hyperparameters. In practice, we follow [4] for implementation details.

C. Experiments

In this section, we evaluate our method on eight diverse tasks. We aim to answer the following three questions through experiments. 1) How does the performance of our method compare to that of the state-of-the-art methods? 2) How well can our method generalize to different configurations and different instances? 3) What enabled our method to generalize to novel instances?

1) Setup: We evaluate our method in both simulation environments and real environments. We use SAPIEN to build simulation environments and conduct experiments [45]. In the real world, our experiment setup is shown in Figure 4. We use ALOHA robot and four RealSense cameras for real-world data collection and testing [49].

Fig. 6: **Similarity Fields Visualization.** We visualize 3D similarity fields in the scene where multiple instances from the same category are presented. It is observed that the similarity fields exhibit similar patterns for different instances. For example, all book titles are highlighted, regardless of whether they are different books in different poses. Furthermore, these highlighted areas represent semantically meaningful parts and are important for various tasks, such as shoelaces, shoe heads, book stands, mug handles, and so on.

We test with various object categories, including shoes, toothbrushes, soda cans, knives, among others. These objects present significant challenges for category-level generalization due to factors such as large geometric variances, ambiguous geometric information, and nuanced geometric details.

2) Tasks: We evaluate various tasks, as shown in Figure 3. Here is a list of task descriptions:

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

3) Comparison with Prior Works: We compare our method with the following baselines:

- **Ours without Semantics**: It takes in the raw point cloud without additional semantic fields as inputs.
- **Diffusion Policy**: This is the vanilla diffusion policy.
- **Diffusion Policy with RGBD**: To compare with the original diffusion policy with the same modality, we add depth observation into the original diffusion policy.

We use success rate as the evaluation metric for different policies. The quantitative result is summarized in Table I. There are several observations we could notice from this table.

First, within seen instances, our method shows no significant difference from the original diffusion policy in the simulation. However, when generalizing to unseen instances, diffusion policy performance degrades obviously. This is because our framework encodes 3D and semantic information about the scene to achieve category-level generalization.To our surprise, the original diffusion policy exhibits significantly worse performance in the real world, even in seen instances. We think that there might be two potential reasons. First, the diffusion policy might be brittle to unnoticeable factors during the deployment, which also shows that our method is robust for real-world deployment. Second, training diffusion policy based on 3D semantic representation could be more sample-efficient.

In addition, our method consistently outperforms ours without semantics, whether for seen instances or unseen instances. This implies that our 3D semantic fields could help policy focus on semantically meaningful parts for task completion.

Furthermore, in situations involving geometric ambiguity and subtle geometric details, the advantage of our method over ours without semantics becomes even more pronounced. Since raw point clouds do not have sufficient information to distinguish geometric ambiguity and pay attention to subtle geometric details, ours without semantics struggles with these tasks. On the contrary, our method shows the ability to effectively disambiguate geometric ambiguity and highlight nuanced geometric information to accomplish the task.

Finally, adding RGBD to the diffusion policy typically results in worse performance. We think that directly adding depth observation to the diffusion policy will make the input space even larger. Therefore, it would require more demonstrations to have sufficient training data coverage,

which makes it less data efficient.

In Figure 3, we see real-world policy rollout results. We notice different failure modes for baseline methods. For diffusion policy and diffusion policy with RGBD, when a novel instance is presented, they may stop early, failing to make progress, either fail to grasp objects, or even lead to unsafe behavior like grasping the blade. Given their reliance on 2D observations as inputs, consequently, they do not possess the capabilities for generalization to unseen instances with varying appearances and geometries.

For ours without semantics, they often fail at objects with geometric ambiguity, like knives, our method could identify the right part for manipulation, such as handles, while ours with raw point fails to differentiate handles and blades. Our method also shows the capability to focus on subtle geometric details. In the can flipping task, by leveraging useful semantic cues, such as the soda can tab, our method successfully placed the can upwards, while ours without semantics may place the soda can upside down.

4) Generalization to Novel Configurations and Instances: We also analyze how well our method can generalize to novel configurations and instances. Figure 5 illustrates the predicted actions of different policies under the same observations. For the seen instances, as shown in the leftmost column, we can see that the predicted trajectories for all policies in the first example point towards the knife handle correctly. However, when the knife blade points towards the right instead of the left, our method without semantics fails to recognize the direction changes and responds accordingly, while the diffusion policy and our method change predicted actions accordingly. This demonstrates that our method is able to understand the semantic difference between knife handle and knife blade, which results in safe robot behaviors.

When the novel instance is presented with similar configurations, our method will predict a successful trajectory guiding towards the knife handle, while diffusion policy will predict a nonsmooth trajectory. Since it takes in 2D observations as inputs instead of 3D and semantic representations, it is brittle to environment factors and lacks category-level generalization capabilities. Although our method without semantics predicts a smooth trajectory, it does not attempt to grasp the knife handle; instead, it approaches the knife blade, which might lead to dangerous actions.

5) Generalization Analysis: We visualize the 3D semantic fields in Figure 6 by overlaying them with the raw point cloud. In the example of shoes, we observe that similar parts of different instances are highlighted, such as shoe heads, shoelaces, and shoe tails. These semantic fields benefit the policy for two primary reasons: 1) Points that receive high similarity scores are identified as semantically meaningful and crucial for task completion. For example, to collect books on the shelf upright, the robot must know where the book top is. Geometric information alone is inadequate for conveying this crucial information for task completion. 2) Semantic fields are consistent across different instances. The book example shows that the activation on the mark is consistent across books with various appearances and

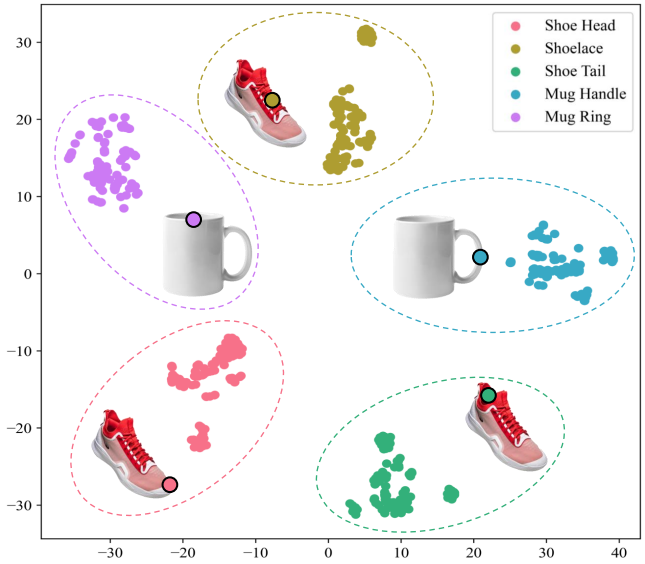


Fig. 7: **t-SNE for Semantic Features with High Similarity.** Points in similarity fields are selected based on the top k similarity scores. We compute their corresponding semantic features and project them into two-dimensional space using t-SNE. It is clearly observed that all features are clustered and separated from each other. Each cluster corresponds to a feature that has semantic meaning, such as mug rings and mug handles.

poses. The category-level consistency enables our method to achieve category-level generalization.

In addition, we also analyze whether our semantic fields could carry desired semantic information, as shown in Figure 7. We initially sample 3D grid points in the workspace and select the top 100 points with the highest similarity score for each semantic field. For the example shown in Figure 7, there will be 500 points in total. Then, we query the descriptor fields and obtain semantic features corresponding to the selected points. We reduce the high-dimensional semantic features to the two-dimensional plane using t-distributed stochastic neighbor embedding (t-SNE). We observe that all points are distinctly separated, and each blob corresponds to one object part. This visualization shows that our semantic fields have a clear semantic meaning for each channel, which helps the policy to reach category-level generalization.