

S²-Diffusion: Generalizing from Instance-level to Category-level Skills in Robot Manipulation

Quantao Yang^{*1}, Michael C. Welle^{*1,2}, Danica Kragic¹, and Olov Andersson¹

Abstract—Recent advances in skill learning has propelled robot manipulation to new heights by enabling it to learn complex manipulation tasks from a practical number of demonstrations. However, these skills are often limited to the particular action, object, and environment *instances* that are shown in the training data, and have trouble transferring to other instances of the same category. In this work we present an open-vocabulary Spatial-Semantic Diffusion policy (S²-Diffusion) which enables generalization from instance-level training data to category-level, enabling skills to be transferable between instances of the same category. We show that functional aspects of skills can be captured via a promptable semantic module combined with a spatial representation. We further propose leveraging depth estimation networks to allow the use of only a single RGB camera. Our approach is evaluated and compared on a diverse number of robot manipulation tasks, both in simulation and in the real world. Our results show that S²-Diffusion is invariant to changes in category-irrelevant factors as well as enables satisfying performance on other instances within the same category, even if it was not trained on that specific instance. Project website: <https://s2-diffusion.github.io>.

Index Terms—Imitation Learning, Learning from Demonstration, Deep Learning in Grasping and Manipulation.

I. INTRODUCTION

IMITATION learning (IL) [1], [2] has shown potential in enabling robotic manipulation in challenging real-world scenarios by learning complex skills from human demonstrations. Still, existing IL methods often struggle to generalize beyond the specific training environments from which the demonstrations are derived. This is an important obstacle as each new environment requires labor-intensive data collection, model fine-tuning, and retraining to adapt the learned policies.

For humans, transferring knowledge between tasks and skills, such as transferring the scooping skill from rice to cereals, is rather straightforward. Scooping rice or cereals may be considered as different instances of the same task for current IL methods. The ability to generalize over such instances is still a challenge and requires rather advanced spatial-semantic understanding [3]. The ability to transfer and

Manuscript received: June, 11, 2025; Revised September, 11, 2025; Accepted October, 4, 2025.

This paper was recommended for publication by Editor Aleksandra Faust upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Knut and Alice Wallenberg Foundation through Wallenberg AI, Autonomous Systems, and Software Program (WASP) and the European Union's Horizon Europe Framework Programme under grant agreement No 101070596 (euROBIN).

*These authors contributed equally.

¹Division of Robotics, Perception and Learning (RPL), KTH Royal Institute of Technology, Sweden. (e-mail: quantao@kth.se).

²Michael C. Welle is also with INCAR Robotics AB, Sweden.

Digital Object Identifier (DOI): see top of this page.

©2026 IEEE

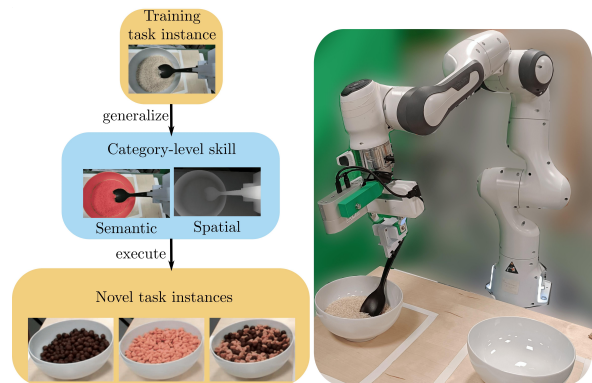


Fig. 1: Our Spatial-Semantic Diffusion policy (S²-Diffusion) not only efficiently completes the task at hand but also enables the generalization of the same skill across diverse contexts and task variations.

generalize over instances removes the necessity for extensive training and also allows for assessing what type of instances one can transfer over - for example, scooping ice cream may be very different from scooping granular material such as cereals or rice. Thus, granular materials may be seen as the same category as rice and cereals, while ice cream is an instance of another category which requires a very different policy when executing the scooping task.

Large pre-trained Visual-Language-Action (VLA) models [4], [5], [6] generalize simple skills such as pick-and-place over a wide range of environments and objects. However, more complex non-prehensile manipulation such as scooping, mug-flipping, cooking shrimp or opening a bottle with a bottle opener stay elusive for such large general models. Recent imitation learning approaches address learning of instances of challenging manipulation tasks [2], [7], [8], [9], [10] but the integration with semantic knowledge in highly-complex manipulation tasks remains a challenge. The aforementioned methods often rely on raw perceptual features and environmental conditions, limiting their applicability to the instances observed during the training. Training a skill via imitation learning, such as a diffusion policy [2], depends on expert demonstrations that often do not cover several instances of the same task. We show that when trained only on a particular instance of a task - such as wiping red scribbles from a whiteboard - the skill fails to transfer already when the scribbles are now green, even if the required action and environment for wiping is exactly the same. This is because the policy did not learn a *whiteboard-wiping* category skill but a single instance of this category namely *red-whiteboard-wiping*.

Motivated by the above challenges, we present a novel approach that integrates spatial prediction with semantic segmentation features from large pre-trained models [11], [12] to generalize from expert demonstrations on a single instance of a task to its *category-level skill* – such as wiping different color scribbles from the whiteboard or scooping different granular material from bowl to bowl. Our method uses a single RGB camera view combined with the proprioceptive information of the robot. As shown in Fig. 1, we extract semantic information using the prompted foundation model [13] and combine it with a depth estimation foundation model [14] to obtain *Spatial-Semantic* observations for the visuomotor diffusion policy learning framework. This allows for invariance to task-irrelevant factors such as background and object textures as well as the capability to generalize from instance-level to category-level skills. Our key contributions are threefold: *i)* We propose to endow visuomotor diffusion policies with spatial-semantic understanding to enable generalization from instance-level to category-level skills in robot manipulation tasks. *ii)* We introduce an efficient representation of the spatial-semantic information via a combination of vision foundation models. The overall framework is real-time viable and requires only a single RGB camera and the robot’s proprioceptive observations. *iii)* Our extensive experiments evaluate the method on a set of robotics manipulation tasks in simulation and the real world, demonstrating the ability to learn generalizable and effective robotic manipulation policies. All real-world experimental videos (246) are on the project website.

II. RELATED WORK

Visual Feature Based Imitation Learning. Visual imitation learning methods [15], [16] have shown strong potential in diverse robot manipulation tasks. While Zhu et al. [3] and Wu et al. [17] highlight the benefits of object-centric and large-scale video pretraining, they do not demonstrate generalization to novel task instances. 3D vision-based methods [18], [19], [20] improve generalization but require costly multi-view RGB-D setups or scene scanning. In contrast, our method uses a semantic-aware representation from a single RGB image. A concurrent work, SAM2Act [21], introduces a multi-view transformer-based robotics policy for improved visual feature representation and multitask generalization, with SAM2Act+ further incorporating a memory-based architecture for episodic recall in spatial memory-dependent manipulation tasks. Closest to our work, Wang et al. [22] integrate explicit spatial-semantic information into policy learning, but rely on multi-view 3D descriptor fields, whereas we achieve similar benefits without extra hardware.

Vision-Language Models in Robotics. Vision-Language Models (VLMs) trained on large internet corpora [23], [24] have become prominent in robotics. Recent works [25], [22] leverage 2D foundation models like CLIP [23] and SAM [26] to construct open-vocabulary 3D representations. Newer models [11], [27] improve open-world generalization. VLMs are also employed to generate high-level task plans [28], [29] or reward functions [30], [31]. Reuss et al. [32] introduce multi-modal diffusion policies for long-horizon tasks. MOKA [33]

Functional goal	flipping				scooping						
Category-skill	mug-flipping	food-flipping		pancake-flipping	bowl-to-bowl	pile-to-container					
Instance-skill	red-mug-flipping	...	blue-mug-flipping	burger-flipping	...	rice-bowl-to-bowl	...	cereal-bowl-to-bowl	sand-pile-to-container	...	gravel-pile-to-container

Fig. 2: Skill abstraction hierarchy for flipping and scooping tasks.

uses keypoint affordances for manipulation, but we enhance generalization further by incorporating both semantic and depth information.

Vision-Language Actions in Robotics. Vision-Language Actions (VLAs) pretrained on large-scale data [34], [5], [35] show strong task reasoning but remain computationally intensive and struggle with fine-grained manipulation. Cross-embodied policies [36] attempt to scale generalization across robots. Instead, our method focuses on compact, efficient policies by leveraging open-vocabulary segmentation masks for policy generalization across manipulation tasks.

III. TOWARDS GENERALIZABLE ROBOTIC SKILLS

What does it mean for a skill to be *generalizable*? Commonly, skills like "mug flipping" learned on a specific instance do not transfer to new instances with different appearance or geometry. Neural policies often overfit to instance-specific features, limiting their applicability. One solution is to explicitly learn features shared across a category. For instance, "mugs" can be characterized by a cylindrical body and handle. Datasets like ShapeNet [37] facilitate learning such category-level features. Object-centric representations work well for defined object classes (e.g., mugs, shoes), however, they struggle with tasks involving materials or actions, such as wiping or scooping. Instead, we advocate an *action-centric* perspective, focusing on *semantic functionality* over object identity.

Skills should be abstracted beyond specific instances. As illustrated in Fig. 2, a functional goal like scooping encompasses various category-level skills and their specific instances. For example, scooping represents transferring granular or semi-solid materials. A category-level skill like bowl-to-bowl scooping covers materials such as rice, cereal, or lentils, whereas instance-level policies (e.g., rice scooping) often fail to generalize. In real-world robotics, demonstrations naturally reflect instance-level setups. Generalization could be attempted by collecting demonstrations across many instances, but this is costly and labor-intensive. Instead, we propose leveraging Spatial-Semantic features extracted from a single instance. By replacing raw RGB inputs with spatial-semantic observations,

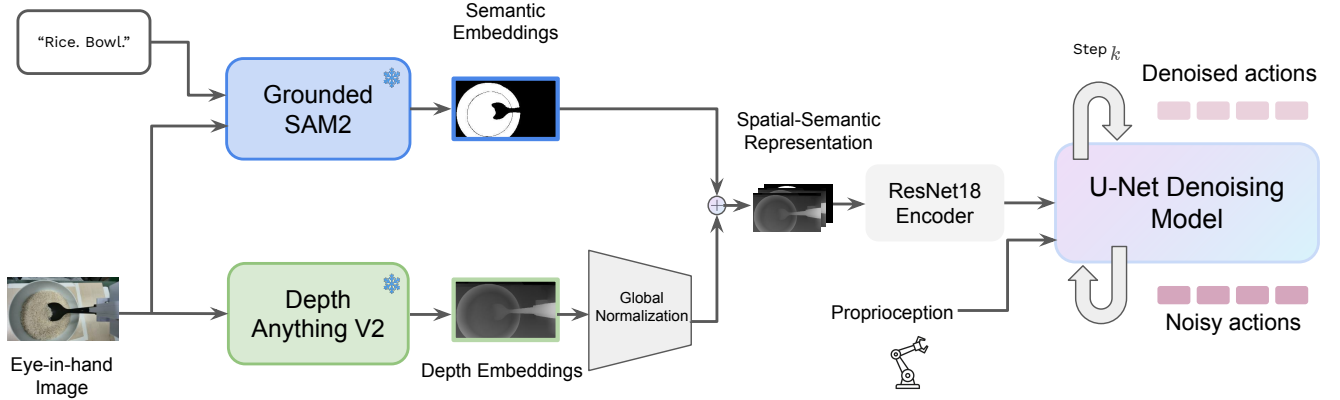


Fig. 3: **S²-Diffusion Architecture**. The architecture is composed of three components: a pretrained semantic segmentation model *Grounded-SAM2* [13], a pretrained depth prediction model *DepthAnythingV2* [12] and a U-Net denoising diffusion policy [2]. We design an object-aware spatial-semantic representation that is leveraged for denoising probabilistic model.

we enable visuomotor diffusion policies to generalize from instance-level training to category-level execution.

IV. PROBLEM FORMULATION

Our goal is to learn a generalizable imitation learning policy by leveraging spatial-semantic representation from pretrained Vision-Language models. We assume access to a dataset $D = \{\tau_i\}_i^N$ of N demonstrated expert trajectories $\tau_i = \{(o_0, a_0), \dots, (o_{T_i}, a_{T_i})\}$ for the task. Our method uses a denoising diffusion process [2] to learn imitation policy π from D . To generalize from individual instances to unseen instances from the same category, we propose to build a spatial-semantic representation. \mathcal{O} is the observation space composed of visual spatial-semantic representation f_v and robot proprioception states q . \mathcal{A} is the action space of robot end-effector commands. We aim to learn a policy $\pi_\theta(a|o) : \mathcal{O} \rightarrow \mathcal{A}$ with parameters θ that predicts action a according to current observation o by leveraging the prior experience contained in the dataset D .

V. METHOD

Our objective is to develop an open-vocabulary spatial-semantic visuomotor policy that can generalize from an individual instance to other unseen instances resulting in a category-level skill. We propose open-vocabulary Spatial-Semantic Diffusion policy (S²-Diffusion), an approach that leverages three main components in policy learning: a semantic segmentation model, a depth prediction model and a diffusion policy shown in Fig. 3. The policy is trained with demonstrations from expert teleoperation, using only RGB images and robot proprioception as the state, and end-effector velocities as the commanded actions respectively.

A. Diffusion for Robot Skill Learning

Following previous works [2], [38], we formulate the visuomotor policy as a conditional Denoising Diffusion Probabilistic Model (DDPM) [39], [40]. Starting from random action a^K sampled from Gaussian noise, the diffusion probabilistic

model ϵ_θ performs K iterations of denoising. This process gradually produces a sequence of actions with decreasing noise levels, a^K, a^{k-1}, \dots, a^0 , until the noise-free action a^0 . Each action denoising iteration is described as:

$$a^{k-1} = \alpha_k(a_k - \gamma_k \epsilon_\theta(a^k, o, k)) + \sigma_k \mathcal{N}(0, I), \quad (1)$$

where o is the observation for the policy. α_k , γ_k and σ_k are referred as noise schedule for each iteration k , and $\mathcal{N}(0, I)$ is the Gaussian noise added to the action.

To learn the action predicting model π_θ , we randomly sample the robot action a^0 from the demonstration dataset D and add the noise ϵ^k for a random iteration k . The training loss for the diffusion model is formulated as:

$$\mathcal{L} = \text{MSE}(a^0, \pi_\theta(a^0 + \epsilon^k, o, k)), \quad (2)$$

where we use an action sampling approach rather than a noise prediction model to enhance the generation of high-dimensional actions.

B. Open-Vocabulary Spatial-Semantic Representation

We aim to design spatial-semantic-aware representation that is leveraged for the observation of the above denoising probabilistic model. We utilize two pretrained VLMs, *Grounded-SAM2* [13] and *DepthAnythingV2* [12], for open-vocabulary semantic segmentation and depth map estimation respectively. We use *Grounded-SAM2* model to perform zero-shot semantic segmentation, leveraging CLIP-based [23] mask classification to segment unseen classes. We combine the extracted features to construct a spatial-semantic representation that is leveraged as the input for the visuomotor diffusion policy. We utilize *Grounded-SAM2* to segment an image into a set of semantic masks (z_1, z_2, \dots, z_n) from visual observations based on text descriptions. We apply pixel-wise maximum pooling for each pixel location (i, j) across all segmentation masks:

$$z_f(i, j) = \max_{s=1}^n z_s(i, j), \quad (3)$$

where $z_s(i, j)$ represents the pixel value at position (i, j) in the s -th mask. In this way, we combine multiple segmentation

Algorithm 1: Learning Open-Vocabulary Spatial-Semantic Diffusion Policy

Input : Semantic query l , image observation o_t , robot state q , *Grounded-SAM2* model C_1 , *DepthAnythingV2* model C_2

- 1 Collect demonstrated trajectories $D = \{\tau_i\}_i^N$
- 2 **for** $epoch\ n=1, N$ **do**
- 3 Sample raw image o_t and robot action a^0
- 4 Obtain spatial-semantic representation
- 5 $z_s = C_1(o_t, l)$, $z_d = C_2(o_t)$
- 6 $z_f(i, j) = \max_{s=1}^n z_s(i, j)$
- 7 $z = z_f \oplus z_d$
- 8 $f_v = \text{ResNet}(z)$
- 9 Add Gaussian noise $\epsilon^k \sim \mathcal{N}(0, I)$ for step k
- 10 $a^k = a^0 + \epsilon^k$
- 11 Train the policy
- 12 $\mathcal{L} = \text{MSE}(a^0, \pi_\theta(a^k, k, f_v, q))$
- 13 **end**
- 14 **return** the trained policy $\pi_\theta(a_t|o_t, q, k)$

masks into a single mask z_f where each pixel represents the most confident prediction from the set of masks.

To improve the spatial reasoning of the imitation policy, we propose to incorporate depth map of the task space into its semantic representation. Specifically, the input image is also processed separately with a pretrained 335M-parameter *DepthAnythingV2* model [12] that shows promising performance in fine-grained details. The model predicts depth estimates z_d relative to the input RGB observations rather than absolute values. This can result in inconsistencies and inaccuracies over extended tasks or manipulation horizons. To address this, we normalize the depth maps during both training and evaluation, ensuring consistency across diverse scenes and improving the model’s robustness and reliability. The resulting semantic and spatial feature vectors are concatenated along the channel dimension to form the spatial-semantic representation $z = z_f \oplus z_d$, where $z \in \mathbb{R}^{(C_s+C_d) \times H \times W}$ with C_s, C_d representing the channel dimensions of the semantic and depth features. Upon generating the spatial-semantic representation, we leverage it as input for a visuomotor diffusion policy [2].

C. Learning Semantic Diffusion Policy

To effectively utilize our open-vocabulary spatial-semantic representations, we adopt the CNN-based diffusion policy architecture [2] as our decision-making backbone. We use Denoising Diffusion Implicit Models (DDIM) [42] as the noise scheduler. In our paper, the observation is composed of spatial-semantic feature f_v and robot proprioceptive state q . The training loss in Equation 2 is defined as:

$$\mathcal{L} = \text{MSE}(a^0, \pi_\theta(\alpha_k a^0 + \beta_k \epsilon^k, k, f_v, q)), \quad (4)$$

where α_k and β_k are used for noise schedule of each step. The learning process for our S²-Diffusion method is described in Algorithm 1. We combine features inferred by VLMs *Grounded-SAM2* and *DepthAnythingV2* to construct the spatial-semantic representation for the action denoising model.

By conditioning on the tuple (a^k, k, f_v, q) , the denoising model learns to predict the clean action by using mean square error loss for action supervision.

VI. EVALUATION

The goals of all our experiments are three-fold:

- 1) to evaluate and compare the performance of our method on challenging robotic manipulation tasks;
- 2) to validate that generalization from instance-level to category-level skill is achieved in a real-world setting;
- 3) to perform ablations in order to investigate the role of the semantic and spatial components.

We first describe the experiments performed in the simulation, followed by the experiments on physical hardware.

A. Simulation Experiments

Experiment Setup. For our simulation experiments, we take advantage of a recent open-sourced large-scale simulation environment, *RoboCasa* [41], which provides expert demonstrations for diverse everyday tasks. We evaluate our method and the baselines on one task of each atomic task category except navigation and group doors and drawers due to their similarity as shown in Fig. 4: *ServeMug*, *CloseDoor*, *TurnOnMicrowave*, *TurnOffFaucet*, *MoveSoda*, and *TurnOnStove* using the provided 50 expert demonstrations from RoboCasa. The prompts used were: "Mug", "Door", "Microwave button", "Faucet", "Soda can", "Knob" respectively.

Baselines. We compare our method with three baseline methods in simulation: 1) **BC-RNN**: a behavior cloning method with recurrent neural network implementation; 2) **BC-Transformer**: a common behavior cloning method with transformer architecture [43]; 3) **Diffusion Policy**: the image-based diffusion policy of [2]; for completeness we also compare against 4) **S²-BC-Transformer**: BC-Transformer trained with spatial-semantic data; and 5) **GenDP**: a multi-view RGB-D method [22]. As RoboCasa does not provide depth, we compare against GenDP on *InsertPencil* and *HangMug* in SAPIEN [44]. In all our experiments, we train for 500 epochs on an NVIDIA RTX 4090 GPU. We set the initial learning rate to $1e-4$, applied a 500-step linear warm-up, and then used a cosine learning rate scheduler for the remainder of training.

Simulation Results. For simulation experiments in RoboCasa, we take advantage of the open-sourced implementation for the baseline methods from *RoboMimic* [43]. In the demonstration datasets for each task, we assume **variations** set by RoboCasa in the color of the target object, as well as differences in the background environment. We evaluate 100 trials with the default seed for each task and the corresponding success rates and the 95% confidence intervals (computed via the Wilson score method for binomial proportions) are listed in Table I. The results are demonstrating matching or statistically significantly superior performance of our S²-Diffusion approach.

Due to the challenging variations in target object and background seen within the RoboCasa dataset, classical imitation learning policies struggle to solve the specific single-stage manipulation task. While the baselines—BC-RNN, BC-Transformer, and Diffusion Policy—show relatively poor performance, particularly on tasks like *ServeMug* and *MoveSoda*,

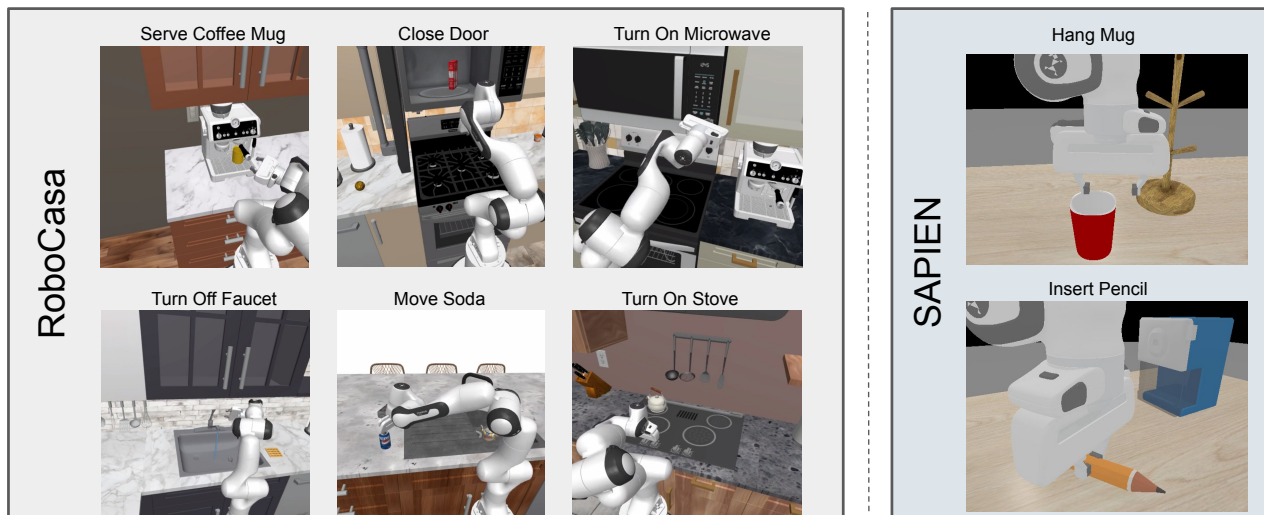


Fig. 4: **Simulated Tasks.** We perform evaluations on six single-stage tasks from a large-scale simulation framework RoboCasa [41]: *ServeMug*, *CloseDoor*, *TurnOnMicrowave*, *TurnOffFaucet*, *MoveSoda*, *TurnOnStove*, and two tasks in SAPIEN simulator: *HangMug*, *InsertPencil*.

TABLE I: Success Rate of Simulation Experiments in RoboCasa

Method	Tasks					
	ServeMug	CloseDoor	TurnOnMicrowave	TurnOffFaucet	MoveSoda	TurnOnStove
BC-RNN	0.02 [0.01, 0.07]	0.03 [0.01, 0.08]	0.05 [0.02, 0.11]	0.27 [0.19, 0.36]	0.06 [0.03, 0.12]	0.26 [0.18, 0.35]
BC-Transformer	0.16 [0.10, 0.24]	0.55 [0.45, 0.64]	0.70 [0.60, 0.78]	0.31 [0.23, 0.41]	0.10 [0.06, 0.17]	0.34 [0.25, 0.44]
Diffusion Policy	0.15 [0.09, 0.23]	0.52 [0.42, 0.62]	0.41 [0.32, 0.51]	0.38 [0.29, 0.48]	0.22 [0.15, 0.31]	0.50 [0.40, 0.60]
S ² -BC-Transformer	0.54 [0.44, 0.63]	0.77 [0.68, 0.84]	0.72 [0.63, 0.80]	0.55 [0.45, 0.64]	0.44 [0.35, 0.54]	0.60 [0.50, 0.69]
S ² -Diffusion (Ours)	0.75 [0.66, 0.82]	0.80 [0.71, 0.87]	0.80 [0.71, 0.87]	0.77 [0.68, 0.84]	0.83 [0.74, 0.89]	0.88 [0.80, 0.93]

TABLE II: Simulation Experiments in SAPIEN

Method	InsertPencil		HangMug	
	Seen	Unseen	Seen	Unseen
GenDP	0.95 [0.76, 0.99]	0.85 [0.64, 0.95]	0.90 [0.70, 0.97]	0.80 [0.58, 0.92]
S ² -Diffusion (Ours)	0.90 [0.70, 0.97]	0.90 [0.70, 0.97]	0.85 [0.64, 0.95]	0.90 [0.70, 0.97]

S²-Diffusion consistently achieves the highest success rates. This lower performance of both BC and diffusion policy on RoboCasa is likely due to the high diversity of scenes in the expert training data. We also trained S²-BC-Transformer on the spatial-semantic dataset. We extended BC-Transformer because in our prior evaluation it exhibited comparable performance with DP. While S²-BC-Transformer achieves competitive results on some tasks, S²-diffusion demonstrates more robust performance. Additionally, S²-BC-Transformer outperforms the RGB-only version, highlighting the value of spatial-semantic representations for improving the generalization of imitation policies. Similar to findings in RoboCasa [41], the image-based DP is sensitive to in object color and background. In contrast, by leveraging spatial-semantic features our method shows better robustness and generality compared to state-of-the-art alternatives. We attribute this to the rich visual representation of combining semantic information and depth estimation for the workspace scene. Finally, Table II shows that our RGB method performs on par with the multi-view RGB-D method GenDP.

B. Real-World Experiments

Experiment setup. We evaluate our S²-Diffusion method on two real-world category tasks: *whiteboard-wiping* and *bowl-to-bowl scooping*. We collected 40 and 60 expert demonstrations for the red-whiteboard-wiping and rice-bowl-to-bowl scooping tasks instances respectively. The demonstrations were obtained by teleoperating a 7-DOF Franka Panda manipulator using the Quest2ROS [45] Oculus app. A single camera providing RGB observation was mounted on the end-effector as well as task-specific tools - such as a sponge for the whiteboard wiping and a spoon for the scooping task. The language prompts for obtaining the semantic observations were "*handwriting, sponge.*" for the whiteboard wiping task and "*rice, bowl.*" for the bowl-to-bowl scooping task.

Baselines and Ablations. We compare our S²-Diffusion method with the visuomotor diffusion method from [2] which uses RGB and proprioceptive information as a baseline, as well as perform a number of ablations on the scooping task - namely a version of our method that only has access to semantic observations (Semantic-Diffusion) and one that only observes the spatial observations (Spatial-Diffusion). Finally, we evaluate generalization to transparent objects.

Real-World Results. The real-world experimental results are reported in Fig. 5. We evaluate our method and the baseline first on the seen environment (red-whiteboard-wiping and rice-bowl-to-bowl-scooping) and evaluate the transferability on other instances of the same category - namely black-

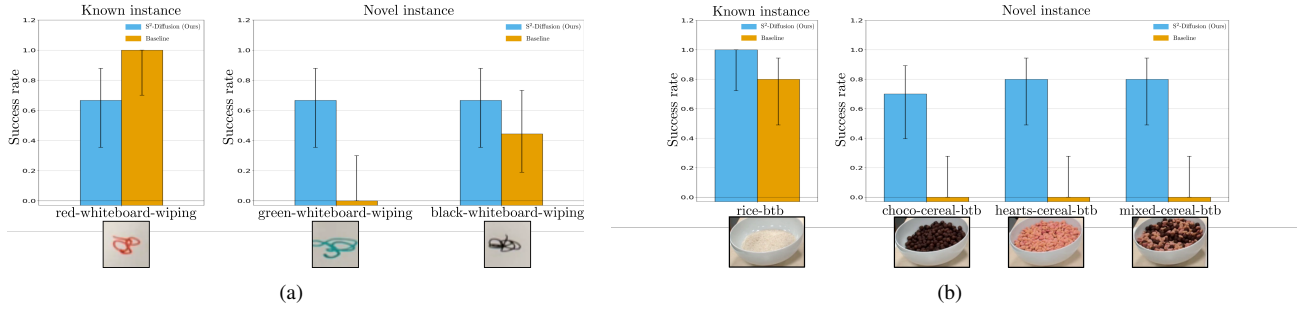


Fig. 5: Comparison of our S^2 -Diffusion and the baseline on two real-world environments: whiteboard wiping and bowl-to-bowl scooping. S^2 -Diffusion and the baseline are trained on *red-whiteboard-wiping* dataset and *rice-bowl-to-bowl-scooping* dataset respectively, then evaluated on the known instances and transferred to unseen instances of the two tasks. Note that for *choco-cereal-btb-scooping*, *hearts-cereal-btb-scooping*, *mixed-cereal-btb-scooping*, and *green-whiteboard-wiping* the baseline diffusion policy shows 0% success rate.

whiteboard- and green-whiteboard-wiping and choco-cereal-, hearts-cereal-, mixed-cereal-bowl-to-bowl-scooping.

Whiteboard-wiping. The task involves wiping scribbles of different colors (red, black, green) from a whiteboard using the end-effector-mounted wet sponge. The results are shown in Fig. 5(a) including the 95% Wilson confidence interval. We define success as the robot completely removing the targeted handwriting using a sponge within 15 seconds. For the task instances that are covered by the training data (red-whiteboard-wiping), the baseline is able to succeed in all nine trials, leading to a 100% success rate, however, once the policy is deployed on different instances of the task such as black-whiteboard-wiping the performance deteriorates to 4/9 (44%) and for the green-whiteboard-wiping task even to 0%. This clearly shows how even small changes in the RGB observations such as changing the color of the scribbles can lead to great deterioration and even complete failure of the skill on this instance - underscoring that the skill learning was indeed on the instance-level and **not** category-level. On the other hand, our S^2 -Diffusion method has consistent performances across all tasks. These results highlight the ability of our S^2 -Diffusion to learn category-level skills for handling novel tasks without requiring additional training or fine-tuning, a capability that is essential for real-world robotic applications. Note that the results of black-whiteboard-wiping are not statistically significant - we attribute this to the fact that the black observations are more similar to red than the green scribbles.

Bowl-to-bowl-scooping. The scooping task is to scoop granular materials from one bowl to another. The amount of material successfully scooped into the target bowl is measured in grams. Task success is defined as the policy’s ability to scoop at least 3 grams of material into the target bowl in under 30 seconds. In Fig. 5(b) the success rates are reported over 10 trials. All real-world experimental videos are available on the project website.

For the in-distribution rice-bowl-to-bowl-scooping task, our S^2 -Diffusion and baseline achieve success rates of 1.0 and 0.8, respectively. However, when transferring the policy to three unseen tasks—choco-cereal-btb-scooping, hearts-cereal-btb-scooping, and mixed-cereal-btb-scooping—the baseline dif-

fusion policy fails entirely, with 0.0 success rate, while our S^2 -Diffusion policy maintains a high success rate of approximately 0.8 by changing the semantic prompt to “*cereal. bowl.*” for all three instances. This demonstrates that our method effectively generalizes from individual instances to unseen other instances of the same category. We can see frames of the baseline diffusion policy and our method in Fig. 6, where the baseline can only succeed on the instance-level task it was trained on and fails to scoop out cereal as the visual observations are too different. Our method on the other hand extracts the semantic mask from the prompt as well as the spatial information via the synthetic depth observation. This leads to the successful execution of the scooping task and to the learning of a successful category-level skill trained on individual instances only using a single RGB observation (same as the baseline) as the original input.

C. Ablation Evaluation

To investigate the significance of integrating spatial-semantic representation, we conduct additional experiments by training the policy using only the semantic representation (Semantic-Diffusion) and only the depth map estimation (Spatial-Diffusion) on the rice and mixed-cereal-bowl-to-bowl-scooping tasks. The results of these experiments are presented in Fig. 7. The full model S^2 -Diffusion, which integrates both modalities, achieves the highest success rates, significantly outperforming the Semantic-Diffusion and also etching out the Spatial-Diffusion versions both on the seen task and also the novel task. When the depth information is excluded, the performance of the policy drops drastically, with success rates falling to 0.4 and 0.5 for the seen and novel tasks, respectively. This significant decline underscores the critical role that spatial information plays in successfully executing 3D tasks. To evaluate generalization to transparent objects, we add a task where it has to close a transparent container. For reference, we also compare against a DP with depth camera input, trained directly on the transparent container. We find that S^2 Diffusion only has a small reduction in success rate while the depth camera policy was not able to complete the task - illustrating the robustness of the S^2 approach in common kitchen tasks involving transparent or reflective surfaces -

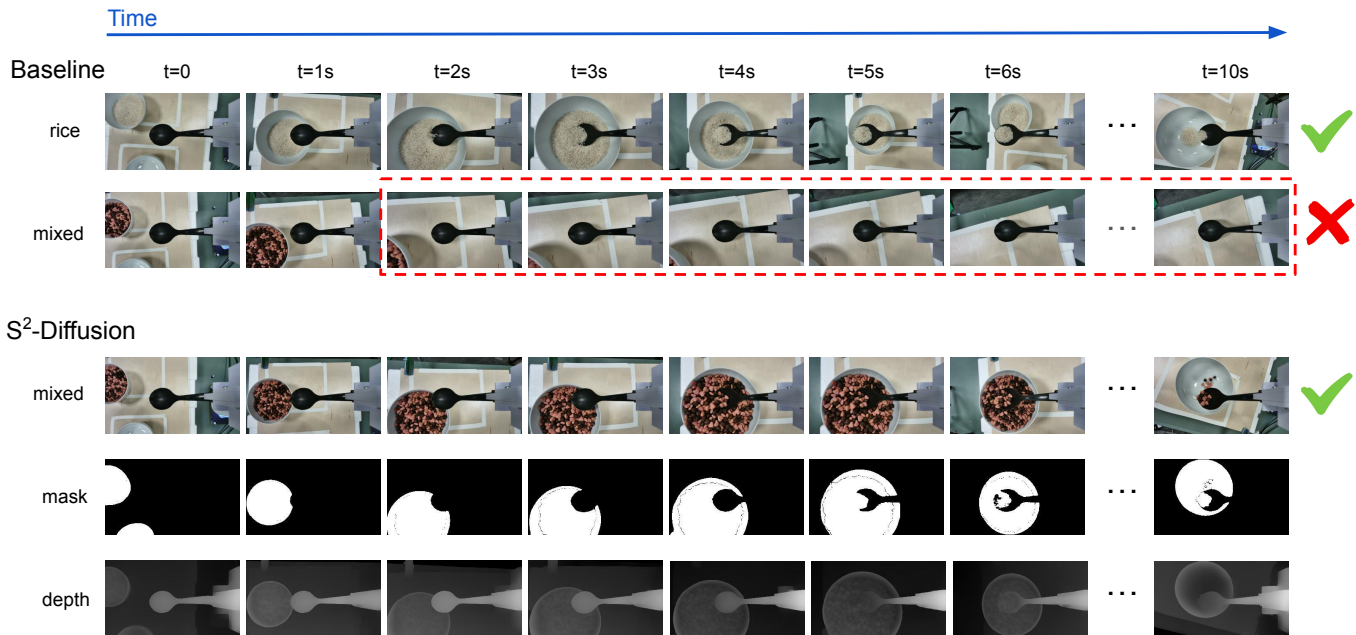


Fig. 6: Frames of the baseline diffusion policy and our S²-Diffusion method for real-world scooping tasks. We show the corresponding semantic mask and synthetic depth for each frame. The baseline can only succeed on the instance-level task it was trained on. Our method succeeds in learning a category-level skill trained on individual instances.

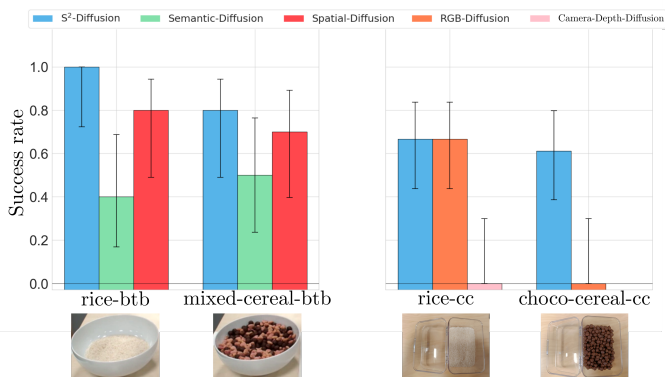


Fig. 7: Ablation study of Semantic and Spatial-diffusion only on a seen task instance (**left**) and an unseen task instance (**right**). S²-Diffusion outperforms models trained with only semantic or spatial observations.

where depth sensors struggle. In contrast, when the mask (semantic representation) is removed, the performance also declines, but to a lesser extent compared to the removal of depth information. While the policy’s ability to generalize across tasks is still reasonable in this case, it benefits from the combined use of both modalities.

VII. LIMITATIONS

As Fig. 2 illustrates, functional goals like *flipping* or *scooping* can be discretized into different category-level skills such as *bowl-to-bowl* and *pile-to-container* scooping. While our method generalizes from instance-level to category-level tasks, it does not generalize well across different categories; for example, a policy trained on *rice-bowl-to-bowl* scooping struggles with a *sand-pile-to-container* task. Although the

abstraction hierarchy in Fig. 2 appears straightforward, real-world functional goals can depend on more complex context and categories like *bowl-to-bowl scooping* may require additional qualifiers. For example, transferring frozen ice cream differs significantly from transferring cereals due to the forces required, although generalization may improve as the ice cream melts. Addressing such nuanced skill ontologies is an avenue for future work. Another limitation is the dependency on the performance of pretrained VLMs as well as proper text prompts for accurate semantic segmentation. Poor semantic or spatial estimates can degrade category-level skill generalization, particularly for greatly different object instances. Fine-tuning VLMs jointly with the policy network is a promising direction for mitigating this dependency.

VIII. CONCLUSION

We introduced S²-Diffusion, an open-vocabulary spatial-semantic diffusion policy that enables generalization from instance-level training data to category-level skills in robotic manipulation. By integrating semantic understanding and spatial representations via vision foundation models, our method learns policies invariant to task-irrelevant visual changes, allowing skill transfer without additional fine-tuning. Through extensive simulations and real-world evaluations, we demonstrated that S²-Diffusion outperforms baselines. In particular:

- **Spatial-Semantic representations enhance generalization:** By combining spatial and semantic information, our method focuses on task-relevant features, enabling instance-to-category transfer.
- **Efficient real-time execution:** S²-Diffusion requires only a single RGB camera, avoiding multi-view setups or depth sensors, making it practical for real-world deployment and attractive for mobile settings.

- **Category-level generalization:** Our evaluations show strong generalization across unseen instances, achieving high performance where baseline policies fail.

Overall, S²-Diffusion represents a step toward enabling robots to generalize skills across variations in objects, materials, and environments, similar to human capabilities.

REFERENCES

- [1] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, “Learning to generalize across long-horizon tasks from human demonstrations,” in *Robotics: Science and Systems (RSS)*, 2020.
- [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.
- [3] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, “Viola: Imitation learning for vision-based manipulation with object proposal priors,” in *Conference on Robot Learning*. PMLR, 2023, pp. 1199–1210.
- [4] A. O’Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models,” *arXiv:2310.08864*, 2023.
- [5] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, “Octo: An open-source generalist robot policy,” *arXiv:2405.12213*, 2024.
- [6] J. Wen, Y. Zhu, J. Li, M. Zhu, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng *et al.*, “Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation,” *arXiv:2409.12514*, 2024.
- [7] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” *arXiv:2304.13705*, 2023.
- [8] T. Gao, S. Nasiriany, H. Liu, Q. Yang, and Y. Zhu, “Prime: Scaffolding manipulation tasks with behavior primitives for data-efficient imitation learning,” *IEEE Robotics and Automation Letters (RA-L)*, 2024.
- [9] S. Fan, Q. Yang, Y. Liu, K. Wu, Z. Che, Q. Liu, and M. Wan, “Diffusion trajectory-guided policy for long-horizon robot manipulation,” *arXiv preprint arXiv:2502.10040*, 2025.
- [10] N. Ingelhart, J. Munkeby, J. van Haastregt, A. Varava, M. C. Welle, and D. Kragic, “A robotic skill learning system built upon diffusion policies and foundation models,” *arXiv:2403.16730*, 2024.
- [11] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, “Open-vocabulary semantic segmentation with mask-adapted clip,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7061–7070.
- [12] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth anything: Unleashing the power of large-scale unlabeled data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10371–10381.
- [13] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, “Grounded sam: Assembling open-world models for diverse visual tasks,” *arXiv:2401.14159*, 2024.
- [14] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, “Depth anything v2,” *arXiv:2406.09414*, 2024.
- [15] O. Mees, L. Hermann, and W. Burgard, “What matters in language conditioned robotic imitation learning over unstructured data,” *IEEE RAL*, vol. 7, no. 4, pp. 11205–11212, 2022.
- [16] H. Ha, P. Florence, and S. Song, “Scaling up and distilling down: Language-guided robot skill acquisition,” in *Conference on Robot Learning*. PMLR, 2023, pp. 3766–3777.
- [17] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, “Unleashing large-scale video generative pre-training for visual robot manipulation,” *arXiv:2312.13139*, 2023.
- [18] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, “3d diffuser actor: Policy diffusion with 3d scene representations,” *arXiv:2402.10885*, 2024.
- [19] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, “Act3d: Infinite resolution action detection transformer for robotic manipulation,” *arXiv:2306.17817*, 2023.
- [20] R.-Z. Qiu, Y. Hu, G. Yang, Y. Song, Y. Fu, J. Ye, J. Mu, R. Yang, N. Atanasov, S. Scherer *et al.*, “Learning generalizable feature fields for mobile manipulation,” *arXiv:2403.07563*, 2024.
- [21] H. Fang, M. Grotz, W. Pumacay, Y. R. Wang, D. Fox, R. Krishna, and J. Duan, “Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation,” *arXiv:2501.18564*, 2025.
- [22] Y. Wang, G. Yin, B. Huang, T. Kelestemur, J. Wang, and Y. Li, “Gendp: 3d semantic fields for category-level generalizable diffusion policy,” in *8th Annual Conference on Robot Learning*, vol. 2, 2024.
- [23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [24] J. Shang, K. Schmeckpeper, B. B. May, M. V. Minniti, T. Kelestemur, D. Watkins, and L. Herlant, “Theia: Distilling diverse vision foundation models for robot learning,” *arXiv:2407.20179*, 2024.
- [25] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiqullah, and L. Pinto, “Ok-robot: What really matters in integrating open-knowledge models for robotics,” *arXiv:2401.12202*, 2024.
- [26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
- [27] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, “Segment everything everywhere all at once,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [28] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv:2204.01691*, 2022.
- [29] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, “Open-vocabulary queryable scene representations for real world planning,” in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11509–11522.
- [30] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, “Eureka: Human-level reward design via coding large language models,” *arXiv:2310.12931*, 2023.
- [31] P. Mahmoudieh, D. Pathak, and T. Darrell, “Zero-shot reward specification via grounded natural language,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 14743–14752.
- [32] M. Reuss, Ö. E. Yağmurlu, F. Wenzel, and R. Lioutikov, “Multimodal diffusion transformer: Learning versatile behavior from multimodal goals,” in *Robotics: Science and Systems*, 2024.
- [33] F. Liu, K. Fang, P. Abbeel, and S. Levine, “Moka: Open-vocabulary robotic manipulation through mark-based visual prompting,” *arXiv:2403.03174*, 2024.
- [34] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Chormanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” *arXiv:2307.15818*, 2023.
- [35] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv:2406.09246*, 2024.
- [36] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, “Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation,” *arXiv:2408.11812*, 2024.
- [37] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Sava, S. Song, H. Su *et al.*, “Shapenet: An information-rich 3d model repository,” *arXiv:1512.03012*, 2015.
- [38] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, “3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations,” in *Robotics: Science and Systems (RSS)*, 2024.
- [39] J. Ho, A. Jain, and P. Abbeel, “Denosing diffusion probabilistic models,” *NeurIPS*, vol. 33, pp. 6840–6851, 2020.
- [40] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann *et al.*, “Imitating human behaviour with diffusion models,” *arXiv:2301.10677*, 2023.
- [41] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, “Robocasa: Large-scale simulation of everyday tasks for generalist robots,” in *RSS*, 2024.
- [42] J. Song, C. Meng, and S. Ermon, “Denosing diffusion implicit models,” *arXiv:2010.02502*, 2020.
- [43] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, “What matters in learning from offline human demonstrations for robot manipulation,” *arXiv:2108.03298*, 2021.
- [44] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, “Sapien: A simulated part-based interactive environment,” in *IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11097–11107.
- [45] M. C. Welle, N. Ingelhart, M. Lippi, M. Wozniak, A. Gasparri, and D. Kragic, “Quest2ros: An app to facilitate teleoperating robots,” in *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions*, 2024.