

PRISM-DP: Spatial Pose-based Observations for Diffusion-Policies via Segmentation, Mesh Generation, and Pose Tracking

Xiatao Sun Yinxing Chen Daniel Rakita

Department of Computer Science
Yale University

{xiatao.sun, j.y.chen, daniel.rakita}@yale.edu

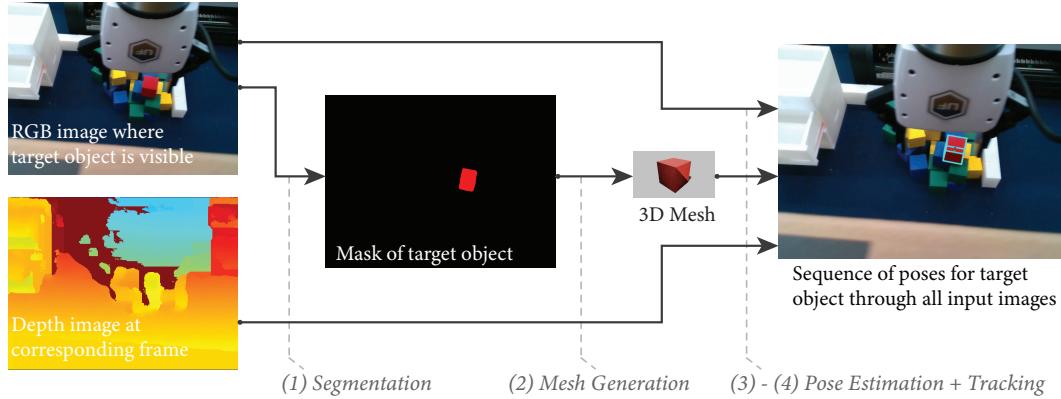


Figure 1: PRISM-DP integrates state-of-the-art learning-based techniques for image segmentation, mesh generation, and unified pose estimation and tracking, enabling efficient diffusion policy learning with estimated poses from RGB-D video streams for simulation or real-world scenarios.

Abstract: Diffusion-based visuomotor policies generate robot motions by learning to denoise action-space trajectories conditioned on observations. These observations are commonly streams of RGB images, whose high dimensionality includes substantial task-irrelevant information, requiring large models to extract relevant patterns. In contrast, using more structured observations, such as the spatial poses (positions and orientations) of key objects over time, enables training more compact policies that can recognize relevant patterns with fewer parameters. However, obtaining accurate object poses in open-set, real-world environments remains challenging. For instance, it is impractical to assume that all relevant objects are equipped with markers, and recent learning-based 6D pose estimation and tracking methods often depend on pre-scanned object meshes, requiring manual reconstruction. In this work, we propose PRISM-DP, an approach that leverages segmentation, mesh generation, pose estimation, and pose tracking models to enable compact diffusion policy learning directly from the spatial poses of task-relevant objects. Crucially, because PRISM-DP uses a mesh generation model, it eliminates the need for manual mesh processing or creation, improving scalability and usability in open-set, real-world environments. Experiments across a range of tasks in both simulation and real-world settings show that PRISM-DP outperforms high-dimensional image-based diffusion policies and achieves performance comparable to policies trained with ground-truth state information. We conclude with a discussion of the broader implications and limitations of our approach.

Keywords: Imitation Learning, 6D Pose Estimation and Tracking, Manipulation

1 Introduction

Diffusion-based visuomotor policies generate robot motions by learning to denoise action-space trajectories conditioned on observations [1]. These observations are commonly sequences of raw RGB images, appealing due to their generality and ease of deployment, requiring just a camera. However, raw image sequences contain substantial amounts of irrelevant information, making it challenging for models to extract meaningful patterns without extensive computational resources, often necessitating hundreds of millions of parameters to fit demonstration data effectively.

In contrast, compact diffusion policies can be trained when using more structured observations, such as the spatial poses (positions and orientations) of key objects over time [1]. Leveraging these representations enables policies to identify relevant patterns with significantly fewer parameters. Yet, obtaining accurate 3D object poses in many scenarios remains challenging, typically requiring external tracking systems or ground-truth annotations, which limits practicality and scalability.

Recent work in learning-based 6D pose estimation and tracking, notably FoundationPose from Wen et al. [2], have introduced promising solutions for accurately estimating and tracking object poses from RGB-D video streams. In theory, these methods could enable compact pose-based diffusion policy training without external tracking setups. However, a significant limitation remains: FoundationPose and similar approaches depend on manually reconstructed or hand-crafted object meshes as input. Prior attempts to integrate FoundationPose into diffusion policy frameworks [3] have been constrained by the need for manual preprocessing of object meshes, hindering scalability in dynamic, open-set environments where object composition varies across tasks.

To address these challenges, we introduce **Pose-tRacking via Image Segmentation and Mesh-generation Diffusion Policies** (PRISM-DP), an approach that leverages segmentation, mesh generation, pose estimation, and pose tracking models to enable compact diffusion policy learning directly from the spatial poses of task-relevant objects. More specifically, PRISM-DP applies a segmentation model to isolate task-relevant objects in input images, it uses the resulting segmentation masks to automatically create 3D triangulated meshes for all target objects using a mesh generation model, then it uses the newly created meshes to initialize a pose estimation and tracking model. Subsequently, the diffusion policy is conditioned on both the robot state and the continuously estimated object poses for the target objects. Crucially, because PRISM-DP uses a mesh generation model, it eliminates the need for manual mesh processing or creation, improving scalability and usability in open-set, real-world environments.

Through several experiments in both simulated and real-world settings, we demonstrate that PRISM-DP facilitates the learning of compact diffusion policies that achieve performance comparable to policies trained using ground-truth object pose information or manually generated meshes, while significantly surpassing image-based diffusion policies with similar parameter counts. To our knowledge, our work is the first to show that diffusion policies conditioned on pose-based observations can consistently outperform those conditioned on raw image observations in a real-world setting, suggesting exciting new directions for future research. Our implementation is publicly available, aiming to foster future research into scalable policy learning through open-set mesh generation and pose estimation.¹

2 Related Works

6D Pose Estimation and Tracking. 6D pose estimation and tracking are foundational technologies in robotics, with decades of extensive study and advancement [4]. Pose estimation refers to the process of determining the 3D position and orientation (collectively known as the 6D pose) of an object from sensor data, such as images or point clouds. Traditionally, pose estimation and tracking have been treated as distinct problems. Pose estimation methods are broadly categorized into model-based approaches, which rely on known CAD or mesh models of the object [5, 6, 7], and model-free

¹<https://github.com/Apollo-Lab-Yale/prism>

approaches, which instead use a few reference images of the target object [8, 9, 10]. Although model-free methods appear more user-friendly, they often require fine-tuning for each novel object [11], suffer from inefficiencies due to internal reconstruction processes [12], or fail under conditions of occlusion and low texture [13]. As a result, model-based methods are generally preferred for achieving robust and reliable pose estimation.

Pose tracking, by contrast, focuses on continuously updating an object’s 6D pose over time by exploiting temporal information from sequential video frames. It is the task of estimating an object’s changing position and orientation across a sequence of frames, typically with an emphasis on efficiency and smoothness. Like pose estimation, tracking methods can be divided into model-free [14, 15] and model-based approaches [16, 17].

Recently, Wen et al. [2] introduced a unified framework called FoundationPose that integrates pose estimation and tracking into a single pipeline, offering improved performance, greater efficiency, and a more streamlined user experience. The prototype implementation of our approach (covered below) uses the FoundationPose framework for pose estimation and tracking, allowing Diffusion Policies to train in the observation-space of key object poses.

Policy Learning with Poses. Pose information provides a compact and structured description of the environment, making it ideal for policy learning. However, deploying pose-based policies in the real world remains challenging due to the difficulty of acquiring accurate 6D poses.

External tracking methods such as AprilTags [18, 19] or Motion Capture [20, 21] are commonly used but require specialized hardware and controlled environments. Others reconstruct scenes into point clouds or voxel maps using depth sensors [22, 23, 24] or LiDAR [25, 26, 27], but such methods are sensitive to environmental changes and require repeated preprocessing if changes occur. Diffusion policies initially trained with ground-truth poses in simulation have been adapted for real-world tasks via visual encoders that process input images [1, 28, 29], but this technique comes at the cost of three to four times increased model size.

Recent work has incorporated FoundationPose [2] into learning-based robotics pipelines [30, 3]. However, Pan et al. [30] use poses only for task planning, while Hsu et al. [3] requires manual mesh scans using consumer devices. In contrast, our approach removes the need for manual reconstruction by integrating segmentation and mesh generation for open-set 6D pose tracking.

3D Generation. The success of large-scale generative models for text and images [31, 32] has recently extended to 3D AI-Generated Content (3D AIGC), with diffusion [33] and Transformer-based architectures [34] driving progress in mesh generation. Recent methods fall into two categories: reconstruction-based approaches [35, 36, 37], which first generate intermediate 3D representations before converting them into meshes, and autoregressive approaches [38, 39, 40], which directly model mesh structures through sequence learning.

Reconstruction-based methods tend to produce higher visual fidelity [35, 36, 37], while autoregressive methods generate meshes with more concise and elegant topology [38, 39, 40] but often at the cost of visual quality [41]. Given the importance of geometric realism for downstream pose estimation and tracking, our approach uses a reconstruction-based mesh generation approach. Specifically, our current prototype implementation (covered below) uses the Meshy model [41, 42].

3 Technical Overview

Our goal is to learn a compact and efficient receding horizon policy π_θ that, given a sequence of past observations $\mathbf{o}_{t-H_o:t}$, predicts a future action sequence $\mathbf{a}_{t:t+H_p}$, where t denotes the current time step, H_o is the observation horizon, and H_p is the prediction horizon.

Commonly, an observation \mathbf{o}_t at a given time t in diffusion policies consists of raw RGB images, $\mathbf{I}_t \in \mathbb{R}^{H \times W \times 3}$. However, to encourage a more compact policy in our approach, we instead reformulate observations \mathbf{o}_t to only include low-dimensional, task-relevant information, e.g., the estimated poses of target objects. Specifically, an observation at time t in our approach is defined as

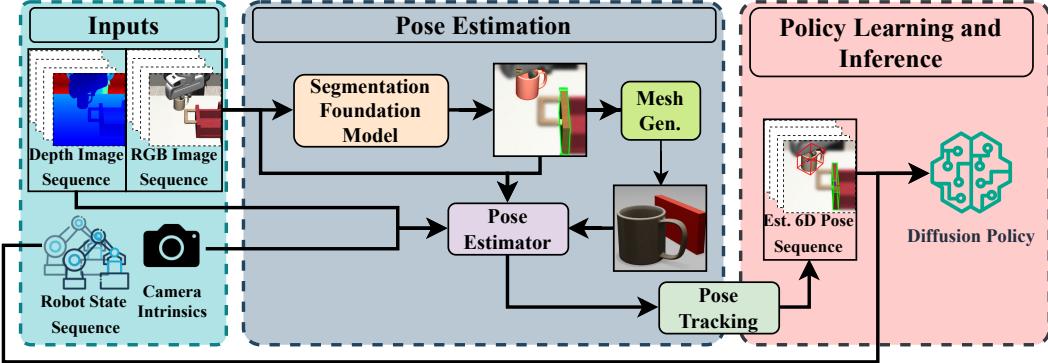


Figure 2: Overview of PRISM-DP. We start from camera intrinsics \mathbf{K} , robot states \mathbf{s}_t , RGB images \mathbf{I}_t and depth images \mathbf{D}_t at all time steps t . Segmentation, Mesh Generation, Pose Estimation, and Pose Tracking models are used to convert these inputs into a sequence of estimate poses for all target objects through a demonstration, which can then be used as observations in a Diffusion Policy.

as $\mathbf{o}_t = \left[\mathbf{s}_t, \left\{ \hat{\mathbf{T}}_i^t \right\}_{i=1}^J \right]$, where $\mathbf{s}_t \in \mathbb{R}^{d_s}$ is the proprioceptive state of the robot, and $\left\{ \hat{\mathbf{T}}_i^t \right\}_{i=1}^J$ is a set of J separate mathematical objects, e.g., $SE(3)$ matrices or unit quaternions, specifying the estimated spatial poses of the J task-relevant objects at the given time t .

The primary challenge through our work lies in inferring the $\hat{\mathbf{T}}_i^t$ poses for all $i \in \{1, \dots, J\}$ and all $t \in \{1, \dots, T\}$, in either simulation or real-world environments, without relying on specialized setups (e.g., markers or motion capture) or requiring an unreasonable amount of manual effort from the user. Our approach, illustrated in Fig. 2, addresses this problem using segmentation, mesh generation, pose estimation, and pose tracking models to conveniently automate the process of converting standard image-based inputs into a sequence of pose estimates for task-relevant objects. The approach takes four inputs: (1) a sequence of RGB images $\mathbf{I}_t \in \mathbb{R}^{H \times W \times 3}$ for all $t \in \{1, \dots, T\}$; (2) a sequence of depth images $\mathbf{D}_t \in \mathbb{R}^{H \times W}$ for all $t \in \{1, \dots, T\}$; (3) a sequence of robot states \mathbf{s}_t for all $t \in \{1, \dots, T\}$; and the camera intrinsics $\mathbf{K} \in \mathbb{R}^{3 \times 3}$.

Using these inputs, our approach follows a five step process:

- (1) For each target object $i \in \{1, \dots, J\}$, the user manually selects a frame index F_i where the i -th target object is first clearly visible in the RGB image \mathbf{I}_{F_i} . If the i -th object is clearly visible at the start of the image sequence, then $F_i = 1$. Our approach currently assumes that once a target object is visible, it will remain in view throughout the rest of the demonstration.
- (2) A segmentation model produces binary masks for each target object at their respective frames of first visibility: $\mathbf{M}_i = \text{Segmentation}(\mathbf{I}_{F_i}), \forall i \in \{1, \dots, J\}$. Each pixel in a mask $\mathbf{M}_i \in \mathbb{R}^{H \times W}$ is 1 if part of the i -th object at time t and 0 otherwise.
- (3) Each mask \mathbf{M}_i is passed to a mesh generation model to obtain a triangular mesh: $\mathcal{M}_i = \text{MeshGen}(\mathbf{I}_{F_i} * \mathbf{M}_i), \forall i \in \{1, \dots, J\}$. The operation $\mathbf{I}_{F_i} * \mathbf{M}_i$ denotes elementwise multiplication, preserving RGB values in \mathbf{I}_{F_i} where the mask is one and setting all other pixels to black (as seen in Fig. 1).
- (4) An approximate pose is calculated at the frames of first visibility for all target objects: $\hat{\mathbf{T}}_i^{F_i} = \text{PoseEstimator}(\mathbf{I}_{F_i}, \mathbf{D}_{F_i}, \mathcal{M}_i, \mathbf{K}), \forall i \in \{1, \dots, J\}$.
- (5) The pose estimates at the frames of first visibility are used to bootstrap a pose tracking model per object: $\hat{\mathbf{T}}_i^t = \text{PoseTracking}(\hat{\mathbf{T}}_i^{t-1}, \mathbf{I}_t, \mathbf{D}_t, \mathcal{M}_i, \mathbf{K}), \forall i \in \{1, \dots, J\}, \forall t \in \{F_i, \dots, T\}$. If an object is not visible at the beginning of a demonstration (i.e., $F_i \neq 1$ for some i), then the object poses at frames preceding F_i are considered null, set as 4×4 matrices of all zeros.

The output of Step 5 is a set of pose estimates over all task-relevant objects and time points. The pose estimates at time t are concatenated with their corresponding robot state, \mathbf{s}_t , to create a sequence of

observations, which are finally used to train a diffusion policy in a compact, low-dimensional space. For interested readers, training model details are found in the Appendix (§A.1).

4 Evaluation

4.1 Prototype System

In this subsection, we provide implementation details for our prototype system that instantiates the approach outlined in §3. For the Segmentation subroutine, the base model in our system is Segment Anything 2 (SAM2) [43]. Task-relevant objects are specified to SAM2 by clicking on the target objects within the sequence of RGB images on their first visible frames, \mathbf{I}_{F_i} . Such inputs are called “point prompts”.

For the MeshGen subroutine, we use the Meshy model [41, 42] due to its ability to produce high-fidelity 3D outputs. Meshy generates a mesh automatically from a masked RGB image, where the region of interest is shown in color and all other regions are blacked out.

Lastly, for the PoseEstimator and PoseTracking subroutines, we use FoundationPose [2], which is able to provide unified pose estimation and tracking.

Importantly, our current system assumes a human-in-the-loop during task execution. Initially, the poses of all task-relevant objects are set to null values (all zeros). When a target object first appears in the field of view, the user pauses execution and provides point prompts to SAM2 through a GUI. FoundationPose then uses the resulting segmentation mask and corresponding mesh object (cached during training) to initialize the object’s pose via pose estimation. Execution then resumes, and FoundationPose continuously updates this object’s pose through tracking. This initialization process is performed separately for each target object.

4.2 Simulation Experiments

Experimental Settings. To evaluate the effectiveness of PRISM-DP, we conduct extensive experiments in simulation across five tasks from Robosuite [44], using a 7-DOF simulated Franka Panda arm. The control is executed in end-effector space, with rotations represented as quaternions. As illustrated in Fig. 3, the tasks are lifting a red cube (Lift), sorting a can into its designated compartment (Can), inserting a square nut onto a square peg (Square), stacking a red cube onto a green cube (Stack), and placing a mug into a drawer for cleanup (Mug).

For training, the datasets for Lift and Can are from Robomimic [45], while datasets for Stack, Square, and Mug are generated using MimicGen [46]. All tasks are trained with 200 demonstration rollouts, except Mug, which uses 300 rollouts to account for its increased complexity.

The simulation evaluation includes a comprehensive set of baselines and architectural variants, as summarized in Table 1. Evaluated methods include Transformer-based (DP-T) and UNet-based (DP-U) diffusion policies trained with ground-truth object poses (GTP), raw RGB images (Img), and estimated 6D poses using ground-truth meshes (GTM). We also include larger image-based variants, DP-T-L Img and DP-U-L Img, that are 8 times and 11 times larger than their standard counterparts (detailed further in the following paragraph). Our proposed method with a Transformer and a UNet is denoted as PRISM-DP-T and PRISM-DP-U.

For policies with image input, the images have resolution of 200×200 . Each UNet-based model uses 24 million parameters in the denoising network, except DP-U-L Img, which uses 261 million. Similarly, all Transformer-based policies use 35 million parameters, with the exception of DP-T-L Img, which scales up to 277 million to better handle high-dimensional image conditions.

All models are implemented in PyTorch [47]. Training and evaluation of policies are performed on a workstation equipped with an AMD PRO 5975WX CPU, dual NVIDIA RTX 4090 GPUs, and 128GB of RAM. Each policy is trained for 200 epochs with batch size of 128, and evaluated over 250 rollouts to calculate the success rate by dividing the total number of evaluation rollouts with number of successful evaluation rollouts.

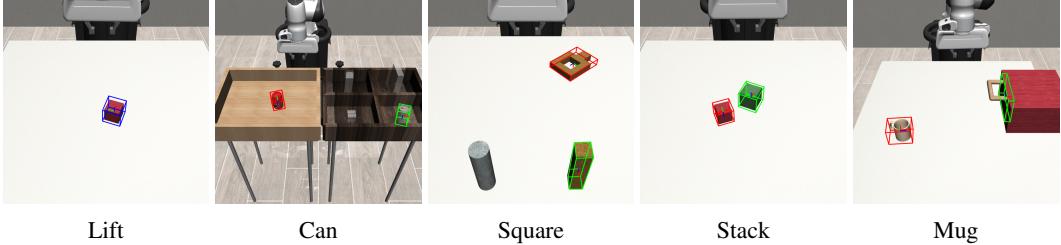


Figure 3: Simulation tasks in Robosuite. Colored 3D bounding boxes indicate objects whose poses are estimated and tracked by FoundationPose.

Table 1: Success rates (SR) and training time per epoch (TE) across five Robosuite tasks. TE is recorded in seconds. Evaluated methods include Transformer-based (DP-T) and UNet-based (DP-U) diffusion policies trained with ground-truth object poses (GTP), raw RGB images (Img), and estimated 6D poses using ground-truth meshes (GTM). We also include larger image-based variants, DP-T-L Img and DP-U-L Img, that are 8 times and 11 times larger than their standard counterparts. Our proposed method with a Transformer and a UNet is denoted as PRISM-DP-T and PRISM-DP-U.

Method	Lift		Can		Stack		Square		Mug	
	SR	TE								
DP-T GTP	0.98	0.89	0.98	1.87	0.91	2.14	0.82	2.88	0.62	9.53
DP-U GTP	0.96	0.87	0.96	2.00	0.81	1.95	0.77	2.53	0.72	8.23
DP-T Img	0.48	5.67	0.25	12.79	0.28	11.99	0.08	16.26	0.27	52.42
DP-U Img	0.48	5.04	0.30	11.27	0.18	10.67	0.26	14.40	0.26	46.70
DP-T-L Img	0.94	9.32	0.90	21.50	0.75	20.30	0.59	27.64	0.58	90.79
DP-U-L Img	0.94	6.69	0.84	15.17	0.74	14.32	0.69	19.47	0.54	63.40
DP-T GTM	0.98	0.85	0.98	1.86	0.86	2.18	0.76	2.87	0.64	9.63
DP-U GTM	0.94	0.87	0.97	1.95	0.79	1.90	0.72	2.65	0.68	8.35
PRISM-DP-T	0.99	0.87	0.98	1.87	0.86	2.16	0.78	2.90	0.61	9.55
PRISM-DP-U	0.95	0.87	0.96	1.99	0.80	1.88	0.72	2.50	0.66	8.33

Results. As shown in Table 1, PRISM-DP consistently outperforms image-conditioned diffusion policies with the same network capacity while having better training efficiency, regardless of whether a Transformer or UNet architecture is used. It achieves comparable performance to diffusion policies trained with poses obtained from ground-truth meshes, indicating that learning-based mesh generation is sufficiently accurate to support downstream pose tracking and policy learning for these tasks. While scaling up the parameter count by 8 times and 11 times in the Transformer-based and UNet-based image models (DP-T-L Img and DP-U-L Img) closes the performance gap for simpler tasks such as Lift and Can, this improvement comes at a significant computational cost and fails to generalize to more challenging tasks. For instance, in tasks like Stack, Square, and Mug, PRISM-DP still outperforms these large image-based baselines, underscoring the value of low-dimensional structured observations.

Notably, PRISM-DP also matches the performance of ground-truth pose-conditioned policies on simpler tasks with only one dynamic object (Lift and Can), but shows a slight performance drop on tasks with multiple dynamic objects (Stack, Square, Mug). This gap is expected, as these baselines have direct access to perfectly accurate object poses, whereas PRISM-DP relies on learned perception modules for object segmentation, mesh generation, and pose estimation.

Analysis on Pose Estimation with Generated Meshes. To further understand the efficacy of PRISM-DP, we evaluate the quality of mesh generation as well as the accuracy of pose estimation and tracking. For mesh generation, we compare meshes produced by the generative model against ground-truth meshes from Robosuite. For pose estimation, we compare the predicted poses, estimated using the generated meshes, to the ground-truth poses transformed into the camera frame.

To assess mesh quality, we render both ground-truth and generated meshes from 30 randomly sampled viewpoints using Blender [48], and compute two commonly used perceptual metrics: Peak Signal-to-Noise Ratio (PSNR) [49] and Fréchet Inception Distance (FID) [50]. PSNR measures pixel-level reconstruction fidelity, where higher values indicate closer alignment, with 20–40 dB considered high quality [51, 52, 53]. FID assesses similarity in feature space, where lower scores are better, and values below 10 typically reflect high-quality generation [54, 51, 52].

To evaluate pose estimation and tracking, we compare the estimated poses, obtained using the generated mesh as input to FoundationPose, with the corresponding ground-truth object poses in the camera frame, across all training samples. Positional accuracy is measured by the Euclidean distance between the predicted and ground-truth translations. Orientation accuracy is measured by the angular distance between predicted and ground-truth quaternions, expressed in radians. We report the mean of each error metric over all samples.

As illustrated in Fig. 4, the generated meshes achieve excellent FID scores, indicating strong perceptual alignment with the ground-truth meshes.

While PSNR scores are in the upper-mid range, suggesting minor pixel-level discrepancies, the low FID underscores that these differences are not semantically significant. This is particularly important given that FoundationPose is a learning-based method that relies on rendered image similarity in feature space for pose refinement. The high perceptual quality of the meshes results in extremely low position errors during pose estimation. Although the orientation error appears comparatively higher, this is largely due to consistent differences in the canonical orientation of the generated meshes versus the ground-truth CAD models. Since FoundationPose tracks pose updates relative to the given mesh geometry, these orientation offsets are typically stable and do not degrade downstream policy performance.

4.3 Real-World Experiments

Experimental Settings. To further validate the proposed approach, we conduct real-world experiments using a dual-arm robotic system (Fig. 5). The setup comprises two 7-DOF xArm7 manipulators, each mounted on a 1-DOF linear actuator. One arm is equipped with a RealSense D435i RGB-D camera for viewpoint control, while the other is fitted with a parallel gripper for manipulation. This configuration enables the policy to dynamically select viewpoints during task execution.

The system operates in a look-at end-effector space: the manipulation arm is controlled via its end-effector’s position and Euler orientation, while the viewpoint arm is controlled by its end-effector position only. The orientation of the viewpoint arm is automatically resolved through an additional IK constraint such that it always looks at the manipulation point on the other arm [55, 56].

We evaluate our approach on two real-world tasks, visualized in Fig. 5. The block stacking task (Block Stack.) requires placing a non-cuboid block onto another, while the drawer interaction task (Drawer Inter.) involves clearing a visual occlusion, retrieving a red cube, inserting it into a drawer, and closing the drawer. The blocks and drawer are 3D printed using meshes from Lee et al. [57] and

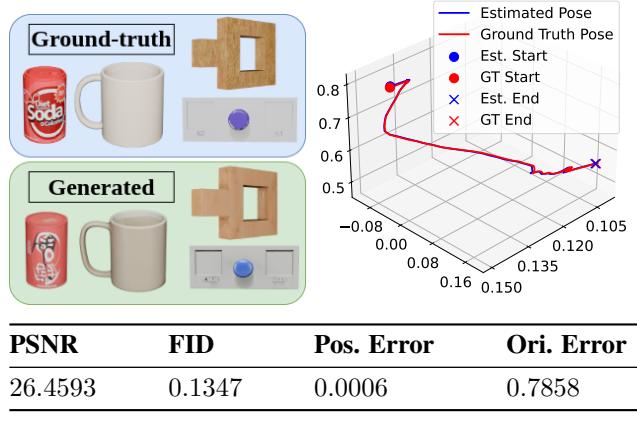
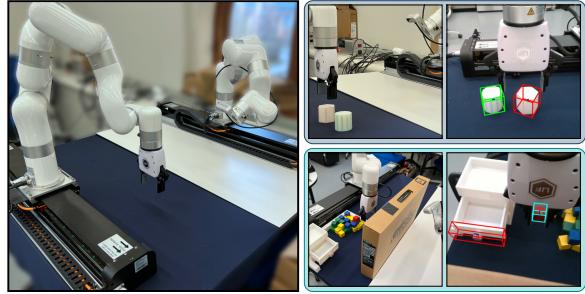


Figure 4: Top left: Examples of ground-truth and generated meshes, shown in the upper and lower rows, respectively. Top right: Visualization of estimated and ground-truth poses of one object during a rollout. Bottom: Average PSNR, FID, and the average positional and orientation errors in radian across all objects in simulation training datasets.

Heo et al. [58], which are also used as ground-truth meshes for evaluation. Each task is trained on a dataset of 200 demonstration rollouts.

The baselines and variants used in real-world evaluation are consistent with those from simulation, excluding baselines that rely on ground-truth poses, which are unavailable in real-world settings. Image-based policies take in images with resolution of 320×240 . For variants that require ground-truth meshes, we use the same meshes employed for 3D printing. The red cube is handcrafted and textured using Blender. All policies are trained for 200 epochs with batch size of 128, and evaluated over 30 rollouts. Success rate is used as the primary evaluation metric. Training and evaluation are conducted on the same workstation used for simulation experiments.

Results. Results for real-world experiments are reported in Fig. 5. Consistent with simulation findings, policies trained using estimated poses from generated meshes perform on par with those using poses from ground-truth meshes, and overall outperform all other baselines in both tasks. In contrast, small-capacity image-based diffusion policies fail completely across both tasks. This likely stems from the increased noise and complexity in real-world RGB observations, which overwhelms their limited representational capacity. Larger image-based diffusion policies achieve marginally lower performance than pose-based policies in the block stacking task, but show a more pronounced performance gap in the drawer interaction task. These results corroborate our simulation findings and reinforce the conclusion that scaling image-based policies to handle complex tasks is increasingly challenging.



Method	Block Stack.		Drawer Inter.	
	SR	TE	SR	TE
DP-T Img	0.00	91.21	0.00	132.40
DP-U Img	0.00	88.13	0.00	126.48
DP-T-L Img	0.87	97.38	0.57	138.13
DP-U-L Img	0.83	93.01	0.63	131.71
DP-T GTM	0.90	7.60	0.83	9.36
DP-U GTM	0.97	6.00	0.80	8.56
PRISM-DP-T	0.93	7.22	0.87	9.52
PRISM-DP-U	0.90	5.91	0.83	8.57

Figure 5: Success rates (SR) and training time per epoch (TE) for all real-world tasks. The evaluated baselines and variants are the same as simulation experiments except for policies learned from ground-truth states since ground-truth states are no longer accessible in real world.

5 Discussion

In this work, we introduce PRISM-DP, a novel diffusion policy framework that integrates segmentation, mesh generation, pose estimation, and pose tracking models to enable efficient policy learning from estimated object poses. By leveraging structured pose-based observations instead of high-dimensional RGB inputs, PRISM-DP significantly improves both task performance and training efficiency. Our experiments demonstrate that PRISM-DP not only outperforms image-based diffusion policies at comparable model scales but also surpasses much larger image-conditioned baselines. These results highlight the importance of structured, low-dimensional representations for scaling up diffusion-based visuomotor policies in robotics.

To our knowledge, this is the first work to demonstrate that diffusion policies conditioned on pose-based observations can consistently and significantly outperform those conditioned on raw image observations in real-world settings on suitable tasks. While our current approach has several limitations that may hinder its immediate broad deployment (outlined below), this initial evidence alone underscores the potential of structured, pose-based observations as a promising direction for future research on diffusion policies or related imitation learning paradigms.

Limitations. Our work has several limitations that suggest directions for future research. First, current mesh generation methods are either too computationally expensive or fail to produce meshes suitable for reliable pose estimation. The Meshy model [42] used in this work, for instance, requires approximately one minute to generate a single mesh. Alternative methods we tested [59, 38, 35] were unable to produce meshes of sufficient fidelity for downstream pose estimation. Although generated meshes often achieve photorealistic appearance, their topology deviates substantially from ground-truth quality. For example, the generated can shown in Fig. 4 contains approximately 1.5×10^5 triangular faces, whereas the corresponding ground-truth mesh has only 956. This excessive triangle count inflates memory usage and increases rendering time during pose estimation and tracking, resulting in latency bottlenecks. Improving the speed, fidelity, and topological quality of mesh generation remains an important direction for future work. Promising approaches may include model distillation, mesh decimation, the use of geometric inductive biases, and incorporation of explicit topological constraints to produce compact yet accurate 3D models that better support efficient pose estimation.

Our current system also assumes that once a target object enters the camera’s field of view, it remains visible throughout task execution. This assumption was not overly restricting in our relatively uncluttered experimental environments but would likely break down in more complex, dynamic settings where occlusions are common. Moreover, our current method for initializing object poses relies on a human-in-the-loop procedure: when a target object first appears, the user must manually pause execution and provide point prompts to SAM2 through a GUI to generate an initial segmentation mask for pose estimation. While effective, this procedure limits scalability to environments with many objects or frequent changes, and reducing this dependence on human input remains an important challenge.

Additionally, our approach is restricted to handling rigid-body motion. It cannot model nonrigid object deformations, such as folding a shirt or manipulating flexible materials—scenarios where, in theory, image-based observation may still succeed at recognizing necessary patterns.

Another limitation stems from our use of unit quaternions to represent the orientation of target objects. While quaternions provide a broadly applicable starting point for reporting on initial results, prior work suggests that alternative rotation representations, such as continuous 6D parameterizations [60], can offer improved learning stability and performance. Exploring more structured orientation parameterizations is an important direction for future improvements.

Overall, while our method demonstrates strong and promising initial performance, addressing these limitations will be critical for scaling to more complex, cluttered, and dynamic real-world environments.

References

- [1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [2] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.
- [3] C.-C. Hsu, B. Wen, J. Xu, Y. Narang, X. Wang, Y. Zhu, J. Biswas, and S. Birchfield. Spot: Se (3) pose trajectory diffusion for object-centric manipulation. *arXiv preprint arXiv:2411.00965*, 2024.
- [4] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He. Deep learning on monocular object pose detection and tracking: A comprehensive overview. *ACM Computing Surveys*, 55(4):1–40, 2022.
- [5] Y. Labb  , L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.
- [6] I. Shugurov, F. Li, B. Busam, and S. Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022.
- [7] D. Chen, J. Li, Z. Wang, and K. Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020.
- [8] M. Cai and I. Reid. Reconstruct locally, localize globally: A model free method for object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3153–3163, 2020.
- [9] F. Li, S. R. Vutukur, H. Yu, I. Shugurov, B. Busam, S. Yang, and S. Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2123–2133, 2023.
- [10] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022.
- [11] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *European Conference on Computer Vision*, pages 298–315. Springer, 2022.
- [12] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022.
- [13] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen. Fs6d: Few-shot 6d pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6814–6824, 2022.
- [14] B. Wen and K. Bekris. Bundlettrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021.

- [15] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023.
- [16] J. Issac, M. Wüthrich, C. G. Cifuentes, J. Bohg, S. Trimpe, and S. Schaal. Depth-based object tracking using a robust gaussian filter. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 608–615. IEEE, 2016.
- [17] M. Stoiber, M. Sundermeyer, and R. Triebel. Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6855–6865, 2022.
- [18] K. T. T. Moyo, J. V. S. Luces, A. A. Ravankar, Y. Hirata, and M. Shota. Enhancing manipulator flexibility: Real-time positional control for variable assembly environments using apriltag markers and edge detection. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 3932–3939. IEEE, 2024.
- [19] A. Angelopoulos, M. Verber, C. McKinney, J. Cahoon, and R. Alterovitz. High-accuracy injection using a mobile manipulation robot for chemistry lab automation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10102–10109. IEEE, 2023.
- [20] P. Zhao, C. X. Lu, B. Wang, N. Trigoni, and A. Markham. 3d motion capture of an unmodified drone with single-chip millimeter wave radar. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5186–5192. IEEE, 2021.
- [21] W. Lindenheim-Locher, A. Świtoński, T. Krzeszowski, G. Paleta, P. Hasiec, H. Josiński, M. Paszkuta, K. Wojciechowski, and J. Rosner. Yolov5 drone detection using multimodal data registered by the vicon system. *Sensors*, 23(14):6396, 2023.
- [22] X. Li, Y. Weng, L. Yi, L. J. Guibas, A. Abbott, S. Song, and H. Wang. Leveraging se (3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in neural information processing systems*, 34:15370–15381, 2021.
- [23] Y. Wu, X. Sun, I. Spasojevic, and V. Kumar. Deep learning for optimization of trajectories for quadrotors. *IEEE Robotics and Automation Letters*, 9(3):2479–2486, 2024.
- [24] J. Choe, S. Im, F. Rameau, M. Kang, and I. S. Kweon. Volumefusion: Deep depth fusion for 3d scene reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16086–16095, 2021.
- [25] W. Wang, B. Wang, P. Zhao, C. Chen, R. Clark, B. Yang, A. Markham, and N. Trigoni. Pointloc: Deep pose regressor for lidar point cloud localization. *IEEE Sensors Journal*, 22(1):959–968, 2021.
- [26] X. Sun, S. Yang, M. Zhou, K. Liu, and R. Mangharam. Mega-dagger: Imitation learning with multiple imperfect experts. *arXiv preprint arXiv:2303.00638*, 2023.
- [27] X. Sun, M. Zhou, Z. Zhuang, S. Yang, J. Betz, and R. Mangharam. A benchmark comparison of imitation learning-based control policies for autonomous racing. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–5. IEEE, 2023.
- [28] X. Sun, S. Yang, Y. Chen, F. Fan, Y. Liang, and D. Rakita. Dynamic rank adjustment in diffusion policies for efficient and flexible training. *arXiv preprint arXiv:2502.03822*, 2025.
- [29] X. Ma, S. Patidar, I. Haughton, and S. James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18081–18090, 2024.

- [30] M. Pan, J. Zhang, T. Wu, Y. Zhao, W. Gao, and H. Dong. Omnimaniip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. *arXiv preprint arXiv:2501.03841*, 2025.
- [31] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [32] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [33] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [35] M. Boss, Z. Huang, A. Vasishta, and V. Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024.
- [36] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.
- [37] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang. Gaussian-dreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6796–6807, 2024.
- [38] Y. Chen, Y. Wang, Y. Luo, Z. Wang, Z. Chen, J. Zhu, C. Zhang, and G. Lin. Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization. *arXiv preprint arXiv:2408.02555*, 2024.
- [39] Y. Siddiqui, A. Alliegro, A. Artemov, T. Tommasi, D. Sirigatti, V. Rosov, A. Dai, and M. Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19615–19625, 2024.
- [40] H. Weng, Y. Wang, T. Zhang, C. Chen, and J. Zhu. Pivotmesh: Generic 3d mesh generation via pivot vertices guidance. *arXiv preprint arXiv:2405.16890*, 2024.
- [41] D. Ebert. 3d arena. <https://huggingface.co/spaces/dylanebert/3d-area>, 2024.
- [42] Meshy LLC. Meshy ai: The ultimate ai 3d model generator for creators, 2025. URL <https://www.meshy.ai/>.
- [43] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [44] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, and Y. Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.
- [45] 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. In *Conference on Robot Learning (CoRL)*, 2021.

- [46] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.
- [47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [48] Blender - a 3d modelling and rendering package. Amsterdam, The Netherlands, 2025. URL <https://www.blender.org>.
- [49] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.
- [50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [52] M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jégou, and J. Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023.
- [53] Y. Chen, A. Janowczyk, and A. Madabhushi. Quantitative assessment of the effects of compression on deep learning in digital pathology image analysis. *JCO clinical cancer informatics*, 4:221–233, 2020.
- [54] T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024.
- [55] D. Rakita, B. Mutlu, and M. Gleicher. An autonomous dynamic camera method for effective remote teleoperation. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 325–333, 2018.
- [56] X. Sun, F. Fan, Y. Chen, and D. Rakita. A comparative study on state-action spaces for learning viewpoint selection and manipulation with diffusion policy. *arXiv preprint arXiv:2409.14615*, 2024.
- [57] A. X. Lee, C. M. Devin, Y. Zhou, T. Lampe, K. Bousmalis, J. T. Springenberg, A. Byravan, A. Abdolmaleki, N. Gileadi, D. Khosid, et al. Beyond pick-and-place: Tackling robotic stacking of diverse shapes. In *5th Annual Conference on Robot Learning*, 2021.
- [58] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *The International Journal of Robotics Research*, page 02783649241304789, 2023.
- [59] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.
- [60] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

A Appendix

A.1 Diffusion Policy Training Details

Conditioned on $\mathbf{o}_{t-H_o:t}$ that includes the estimated poses and robot states, the policy π_θ is implemented as a diffusion policy that outputs a predicted sequence of actions $\mathbf{a}_{t:t+H_p}$. For a diffusion model in general [33], due to Markov property, the forward noising process can be reparameterized using a closed-form solution that directly samples \mathbf{x}_k from \mathbf{x}_0

$$\mathbf{x}_k = \sqrt{\bar{\alpha}_k} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_k} \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$$

where $\boldsymbol{\varepsilon}$ is a standard Gaussian noise vector sampled with mean 0 and variance 1, $\bar{\alpha}_k$ is a variance scheduler parameter of the noise scheduler, and k is the timestep for denoising or introduce noise.

Intuitively, $\bar{\alpha}_k$ controls how much of the original clean signal \mathbf{x}_0 is preserved versus how much noise is introduced at timestep k . As more steps of noise are added by increasing k , $\bar{\alpha}_k$ generally decreases, causing \mathbf{x}_k to become progressively noisier.

During inference, the goal is to reverse this forward diffusion process. Instead of recovering \mathbf{x}_0 directly in a single step, the model learns to gradually denoise \mathbf{x}_k through an iterative procedure. Specifically, at each timestep k , the model predicts the noise component $\boldsymbol{\varepsilon}$ present in the current noisy sample \mathbf{x}_k , using a learned denoising network $\epsilon_\theta(\mathbf{x}_k, k, \mathbf{y})$, where \mathbf{y} represents conditioning information such as observations.

Given this predicted noise, the model constructs the mean $\mu_\theta(\mathbf{x}_k, k, \mathbf{y})$ of the reverse Gaussian distribution $p_\theta(\mathbf{x}_{k-1} | \mathbf{x}_k, \mathbf{y})$ used to sample the denoised version \mathbf{x}_{k-1} at the previous timestep:

$$\mu_\theta(\mathbf{x}_k, k, \mathbf{y}) = \frac{1}{\sqrt{\alpha_k}} \left(\mathbf{x}_k - \frac{1 - \alpha_k}{\sqrt{1 - \bar{\alpha}_k}} \cdot \epsilon_\theta(\mathbf{x}_k, k, \mathbf{y}) \right)$$

where α_k is another variance-related parameter derived from the noise schedule. Specifically, $\alpha_k = 1 - \beta_k$ where β_k is the variance increment at step k .

The model ϵ_θ is trained to predict $\boldsymbol{\varepsilon}$ given \mathbf{x}_k , the denoising timestep k , and the conditioning \mathbf{y} by minimizing the mean squared error loss between the predicted noise by ϵ_θ and the ground-truth noise $\boldsymbol{\varepsilon}$

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\varepsilon}, k} \left\| \boldsymbol{\varepsilon} - \epsilon_\theta(\mathbf{x}_k, k, \mathbf{y}) \right\|^2$$

In PRISM-DP, we instantiate this diffusion model ϵ_θ as our policy π_θ on action sequences, replacing \mathbf{x}_k with $\mathbf{a}_{t:t+H_p,k}$, and conditioning on the observation window $\mathbf{y} = \mathbf{o}_{t-H_o:t}$. Since each observation concatenates a robot state and object poses, the training objective becomes:

$$\mathcal{L}_{\text{PRISM}}(\theta) = \mathbb{E}_{\mathbf{a}_{t:t+H_p,0}, \boldsymbol{\varepsilon}, k} \left\| \boldsymbol{\varepsilon} - \epsilon_\theta(\mathbf{a}_{t:t+H_p,k}, k, [\mathbf{s}_t, \{\hat{\mathbf{T}}_i^t\}_{i=1}^J]) \right\|^2.$$