

## A Ablation Study

### A.1 Geometric and Spatial (GS) Representation

Direct grasp attempts in dense cluttered scenes often result in unsafe collisions with surrounding non-target objects, we therefore introduced a Geometric and Spatial Representation to encourage collision-minimized behavior. At each time step, 200 points are sampled from the target object and 50 points from surrounding objects using their mesh files. For each finger joint, distances to the five nearest sampled points from both target ( $d_{hand}$ ) and non-target objects ( $d_{neg}$ ) are computed to form the representation, as shown in Fig. 6. This representation is embedded into observation space and reward function, providing essential spatial and geometric information for clutter-aware grasping, while bypassing the need for time-consuming point-cloud rendering.

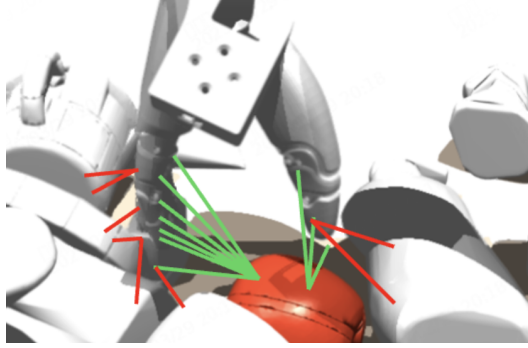


Figure 6: Visualization of Geometric and Spatial Representation. The distances to the five nearest sampled points from both target ( $d_{hand}$ ) and non-target objects ( $d_{neg}$ ) are labeled in green and red, respectively.

To better understand the contribution of GS-representation, we introduce two ablation experiments: (1) one without the proposed geometry-and-spatial representation (w/o GS-repr) (2) and another without the negative object representation (w/o negative repr). In the w/o GS-repr setting, we replaced the GS-representation with a simplified distance vector: a distance vector from the object center to the hand palm, the observation includes both positive and negative object distance vectors but lacks object geometric information. In the w/o negative repr setting, we used the same single-object training observations as in our method, but during the clutter grasping stage, the negative object components were masked out from the input representation. Figure 7 shows that while the success rates of these methods are comparable, our method achieves significantly lower maximum contact force, indicating that it learns to avoid collisions with surrounding objects when grasping the target. The qualitative difference is more apparent in rollout visualizations (Fig. 8). Both ablated variants w/o GS-repr and w/o negative repr frequently attempt direct top-down grasps, ignoring clutter and applying excessive force on clutter especially when the target is partially occluded leading to unsafe behaviors. In contrast, the GS-based policy exhibits more strategic behavior: gently repositioning occluding objects from the side and approaching the target from angles that minimize collision. Such human-like strategies are critical for sim-to-real transfer. Importantly, this also leads to a smooth transition into the safety curriculum stage, with negligible performance drop (Table 2).

### A.2 Clutter-Density Curriculum Learning

Learning dexterous grasping in cluttered scenes purely through random trial-and-error is extremely challenging. As shown in Figure 9, a policy trained from scratch under identical settings fails to make progress, with success rates remaining at zero. To address this, we introduce a two-stage Clutter-Density Curriculum: the policy is first trained on general single-object grasping, then fine-tuned in cluttered scenes to develop strategic, human-like behaviors. This staged approach enables effective learning, as illustrated by the performance curve in Figure 9.

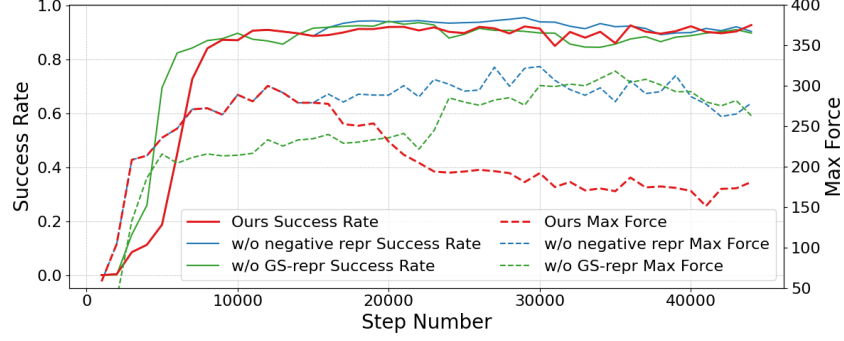


Figure 7: Learning curves of the cluttered-scene policies with or without the Geometric and Spatial Representation we introduced.

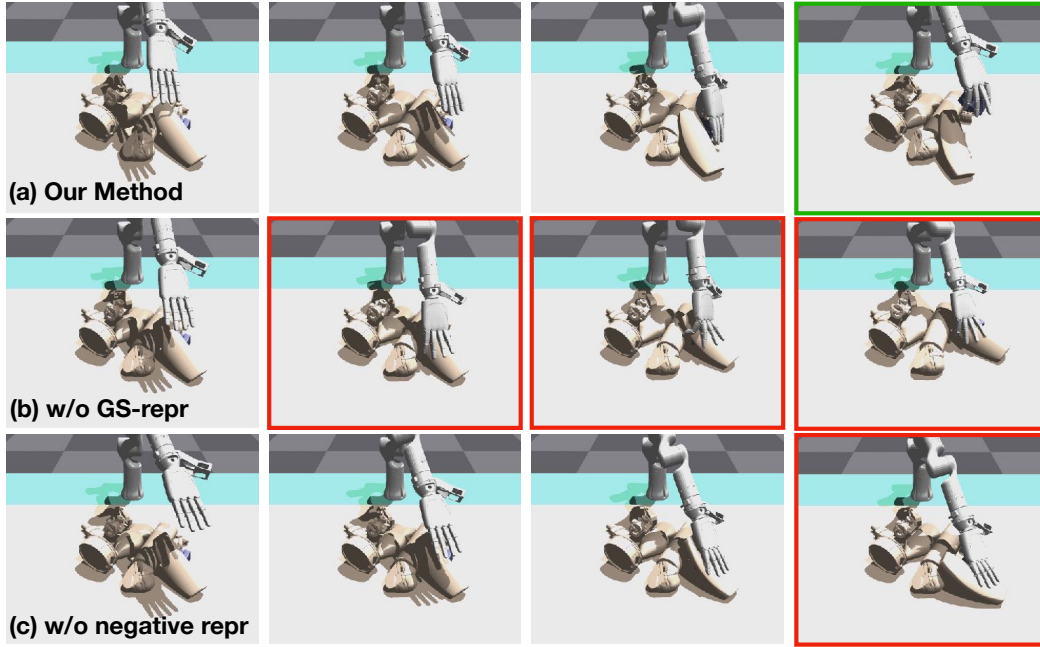


Figure 8: Cluttered Scene Policy Strategy Comparison: (a) Policy trained with GS-Representation, (b) Policy trained without GS-Representation, (c) Policy trained without negative representation. Green bounding boxes indicate successful grasps, while red bounding boxes highlight unsafe or risky actions.

## 500 B Failure Case Analysis

501 Our policy struggles with grasping extreme shapes and sizes due to the limitations of hand morphol-  
 502 ogy. Specifically, the policy fails with large objects or excessively flat ones. Additionally, due to the  
 503 limited working space of our robotic arm, we did not strictly constrain the scene generation to the  
 504 arm’s operational range. As a result, some failures occur when objects fall outside the working area  
 505 or are pushed out of the range by the arm during grasping. Check our website for failure videos.

## 506 C Data Scaling for Scene-Level Generalization

507 In this section, we investigate how the performance of student policy scales with the number of  
 508 collected trajectories. Specifically, we use a consistent teacher policy to collect trajectories in iden-  
 509 tical scenes, but employ varying scales of trajectory data for student policy distillation. Consistent

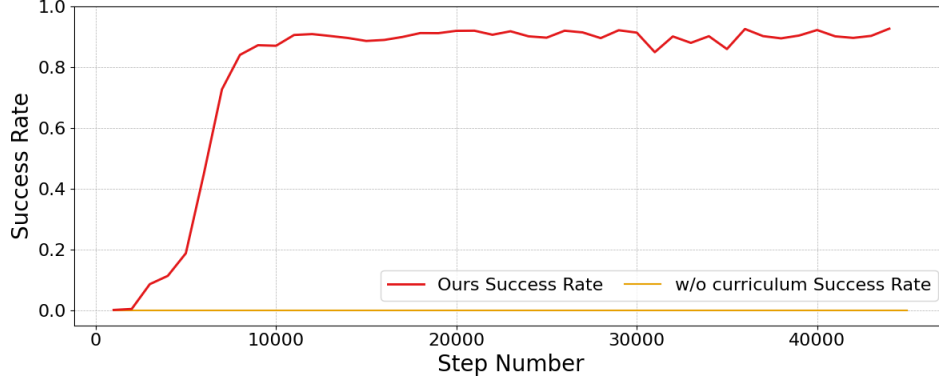


Figure 9: Learning curves of the cluttered-scene policies (1) trained from scratch directly in Cluttered Scene, (2) initialized with stage 1 general single-object grasping policy.

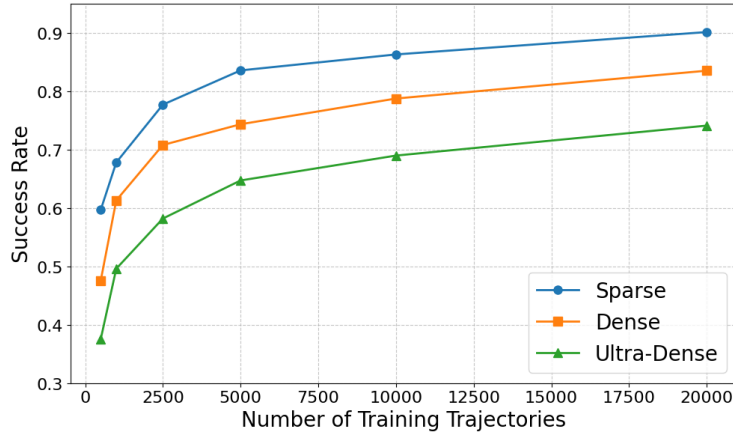


Figure 10: Data Scaling for Generalization

with the main manuscript, we conduct evaluations across three scenario types: Sparse, Dense, and Ultra-Dense. As illustrated in Figure 10, results from these representative scenes consistently show significant performance improvements as the data scale increases. This enhancement in performance is attributed to the richer diversity and more comprehensive coverage of the state-action space afforded by larger datasets, which facilitates more effective learning and generalization.

## D Implementation Details of Teacher Policy Learning

### D.1 Teacher Coarse-to-Fine Data-Efficient Learning

The high degrees of freedom (DoFs) in dexterous hands pose significant challenges for efficient learning. To address this, we model the grasping process as a coarse-to-fine trajectory: a coarse approach phase followed by a fine interaction phase. During the approach, hand DoFs are frozen by masking hand actions with their initial state, leveraging privileged object-hand distance information. Once the hand is sufficiently close to the object, the full action space is enabled.

### D.2 Teacher Observation Space

The observation space for teacher  $o_t \in \mathcal{O} = \mathbf{J}^a \times \mathbf{J}^h \times \mathcal{O}^E$  consists of the representation introduced  $d_{hand}, d_{neg}$ , the joint state, the state of the object, the distance from the table.

Table 4: Teacher observation space

Index	Description
0 – 19	DoF positions (unscaled) $\mathbf{J}^a \in \mathbb{R}^7, \mathbf{J}^h \in \mathbb{R}^{12}$
19 – 22	End-effector position (XYZ)
22 – 25	End-effector orientation (Euler angles: roll, pitch, yaw)
25 – 28	End-effector linear velocity
28 – 31	End-effector angular velocity
31 – 34	Vector from object to middle point
34 – 37	Middle point position
37 – 44	Object pose (position + quaternion)
44 – 47	Object linear velocity
47 – 50	Object angular velocity
50 – 57	Object goal pose (position + quaternion)
57 – 90	Distance features (flattened)
90 – 123	Negative distance features (flattened)
123 – 128	Safety: finger-tip-to-table height (5 values)

### 525 D.3 Teacher Reward Function

526 At each timestep, the complete reward function is defined as:

$$r = (c_1 \cdot r_{grasp} + c_2 \cdot |d_{pos}|) \cdot (1 - \bar{d}_{neg}^{\max}) \quad (5)$$

527 Here,  $\bar{d}_{neg}^{\max}$  denotes the maximum absolute distance to any negative object observed over the entire  
 528 grasping episode. It serves as a global penalty term to discourage risky proximity to non-target  
 529 objects at any point during execution, and  $r_{grasp}$  is defined as:

$$\begin{aligned}
 r_{grasp} = c_5 \cdot (c_6 \cdot (0.2 - \|p_{current} - p_{goal}\|_2) & \quad \text{goal distance reward} \\
 + c_7 \cdot |d_{hand}| & \quad \text{reach reward} \\
 + c_8 \cdot |d_{mid}|) & \quad \text{middle point reward}
 \end{aligned} \quad (6)$$

530 where  $c_1, c_2, c_5, c_6, c_7, c_8 > 0$  are weighting coefficients.

531 **Stage 1** For general single-object grasping, the representation-related components in both obser-  
 532 vation and reward are zero-padded, with reward function being:

$$r_{stage1} = c_1 \cdot r_{grasp} \quad (7)$$

533 **Stage 2** For strategic cluttered-scene grasping, with reward function is:

$$r_{stage2} = r \quad (8)$$

534 **Stage 3** For safety finetuning, the reward function for the curriculum is:

$$r_{stage2} = r + r_{safe} \quad (9)$$

### 535 D.4 Safety Curriculum

536 The safety threshold starts at an initial value of  $f_{threshold} = 200$  and gradually decreases to a final  
 537 value of  $\bar{f} = 50$ . After reaching a certain success rate threshold, the safety threshold is reduced  
 538 by 5 units at each step. During the actual training, we disable collisions between the hand and the  
 539 table, and instead introduce a penalty term based on the distance between the fingertips and the table  
 540 surface.

### 541 D.5 Training Details

542 All the training and experiment in the paper were run on a single GeForce RTX 4090 GPU with  
 543 i9-13900K CPU.



---

**Algorithm 1** Safety Curriculum

---

```
1: Initialize empty FIFO queue  $Q$  of size  $K$ ,  $\Delta T = 0$ , safety threshold  $\lambda = \lambda_0$ 
2: for  $i \leftarrow 1$  to  $M$  do
3:    $\tau = \text{rollout\_policy}(\pi_\theta)$  ▷ get rollout trajectory
4:    $\pi_\theta = \text{optimize\_policy}(\pi_\theta, \tau)$  ▷ update policy
5:    $\Delta T = \Delta T + 1$ 
6:   if  $i \bmod L = 0$  then
7:      $w = \text{evaluate\_policy}(\pi_\theta)$  ▷ get success rate  $w$ 
8:     append  $w$  to the queue  $Q$ 
9:     if  $\text{avg}(Q) > \bar{w}$  and  $\Delta T > \Delta T_{\min}$  then
10:       $\lambda = \min(\lambda + \Delta\lambda, \lambda_{\max})$  ▷ tighten safety constraint
11:       $\Delta T = 0$ 
12:    end if
13:  end if
14: end for
```

---

Hyperparameters	Value
Num mini-batches	4096
Num opt-epochs	5
Num episode-length	8
Hidden size	[1024, 512, 256]
Clip range	0.2
Max grad norm	1
Learning rate	3e-4
Discount ( $\gamma$ )	0.99
GAE lambda ( $\lambda$ )	0.95
Init noise std	—
Desired kl	0.02
Ent-coef	0.0

Table 5: Hyperparameters of PPO.

## 544 E Implementation Details of Student Policy Distillation

545 We offline distilled the teacher policy  $\pi^E$  trained with privilege state information  $\mathcal{O}^E$  into a student  
546 policy  $\pi^S$  that takes sensory observations  $\mathcal{O}^S \in \mathbb{R}^{4 \times 109}$  that can be obtained in the real world. The  
547 student observation space contains the robot joint position  $\mathcal{O}_{robot} \in \mathbb{R}^{13}$  and the partial point cloud  
548  $\mathcal{O}_{pc} \in \mathbb{R}^{4096}$  from a fixed side-view camera cropped and transformed to the robot frame.

### 549 E.1 Student Observation

550 The point cloud observation  $\mathcal{O}^{pc} = \mathcal{O}^p \times \mathcal{O}^g \times \mathcal{O}^s$  is composed of three components: (1)  $\mathcal{O}^p \in$   
551  $\mathbb{R}^{4 \times 3584}$ , the partial point cloud observation from a single third-person camera, with a 1D mask to  
552 indicate the grasping target; (2)  $\mathcal{O}^g \in \mathbb{R}^{4 \times 512}$ , a synthetic ground point cloud replacing the table  
553 surface to mitigate sensor noise; and (3)  $\mathcal{O}^s \in \mathbb{R}^{4 \times 1024}$ , the synthetic robot point cloud observation  
554 (Sec.4.2.2) with a 1D mask to differentiate between real and synthetic point clouds.

### 555 E.2 Multi-Level Clutter Density Dataset Generation

556 we collect a dataset  $\mathcal{D}^E$  consisting of 20,000 successful trajectories generated by the teacher pol-  
557 icy  $\pi^E$  across 500 clutter scenes with varying levels of clutter density. Demonstrations are evenly  
558 distributed across the different clutter levels. The student policy is trained using Imitation Learning  
559 (IL) with the DP3 algorithm[33]. where a random batch size of 120 is sampled from the dataset.  
560 Observations are normalized based on the data distribution’s statistics.

### 561 E.3 Point Cloud Observation Processing

562 The point cloud is created from a single-view depth image. The point-cloud pre-processing includes  
 563 four steps: (i) cropping the point cloud to the workspace region using a manually defined bounding  
 564 box, (ii) downsampling the point cloud to 3584 points, (iii) transforming the point cloud from the  
 565 camera frame to the robot base frame, and (iv) replacing the table surface with a synthetic point cloud  
 566 (512 points) to mitigate point cloud holes caused by the flat table. However, we found that adding  
 567 Gaussian noise and random transformation to point cloud during DP3 training did not improve policy  
 568 performance in real-world settings.

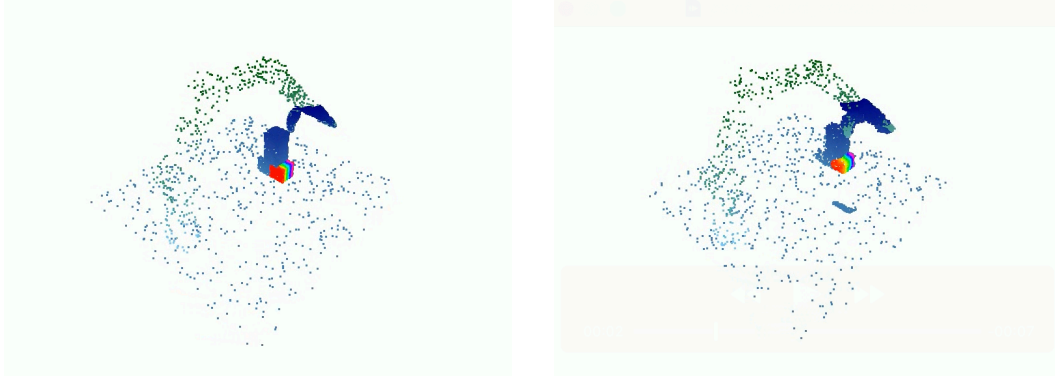


Figure 11: Point-cloud comparison between simulation(Left) and real world(Right).

### 569 E.4 Training Details

Table 6: Key Hyperparameters of SimpleDP3 Policy.

Hyperparameters	Value
Downsample dims	[128, 256, 384]
Encoder output dim	64
Crop shape	[80, 80]
Horizon	4
Num observation steps	2
Num action steps	1
Kernel size	5
Num groups	8
Diffusion steps (training)	100
Diffusion steps (inference)	10
Learning rate	1e-4
Optimizer	AdamW
Weight decay	1e-6
Betas	(0.95, 0.999)
EMA power	0.75
EMA max value	0.9999
LR scheduler	cosine
LR warmup steps	500

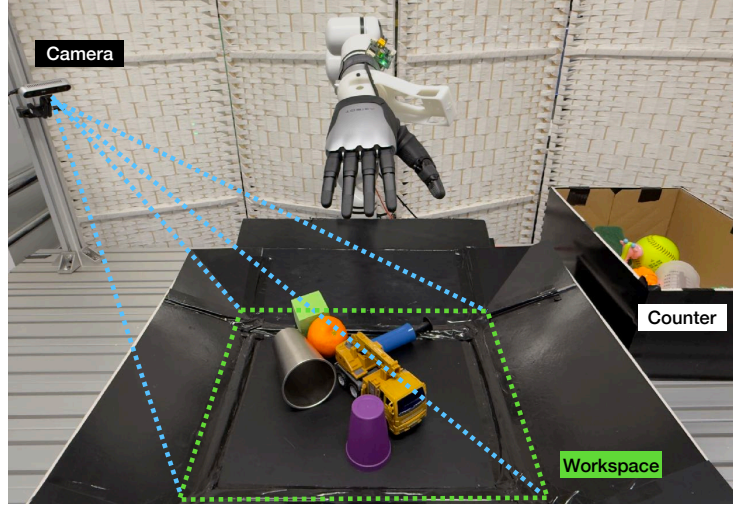


Figure 12: Real-World Setup

## 570 **F Implementation Details in Real-World**

### 571 **F.1 Real-World Hardware Setup**

572 For real-world deployment, the target object is selected and segmented using SAM2 [60]. The  
 573 resulting binary mask is projected onto the point cloud using one-hot encoding to isolate the target  
 574 object. The entire system operates at a frequency of 15 Hz.