# Attention-DP3: Spatially Object-aware 3D Diffusion Policy via Geometry-aligned Attentional Conditioning

Changbo Yan*, Zhongbo Zhang*, Zaibin Zhang*, Lijun Wang, Yifan Wang✉, Huchuan Lu

Dalian University of Technology,
wyfan@dlut.edu.cn
* Equal contribution ✉ Corresponding author

**Abstract.** 3D point-cloud observations are inherently ambiguous in complex, cluttered manipulation scenes, where target objects may be partially occluded or tightly intermingled with visually similar distractors. As a result, standard 3D diffusion policies often struggle to localize and exploit task-relevant geometry as scene complexity grows. We propose **Attention-DP3**, a spatially object-aware 3D diffusion policy that injects object-level geometric cues via attention while keeping the DP3 diffusion backbone unchanged. Our pipeline performs open-vocabulary 2D segmentation on RGB images, then lifts predicted target masks into 3D using calibrated camera geometry to obtain object-centric geometric priors. We incorporate these cues through Tri-field Attentional Conditioning, which constructs three complementary fields: (i) a targetness field to anchor the target object, (ii) an intra-target saliency field to emphasize task-relevant geometry within the target, and (iii) a backgroundness field to suppress distractors and clutter. Experiments on Adroit, DexArt, MetaWorld, and the real-world SO101 platform show consistent improvements over DP3, achieving state-of-the-art performance across benchmarks. Notably, as distractor objects increase, DP3 drops sharply, whereas Attention-DP3 remains stable and outperforms DP3 by up to 31% under heavy clutter. The code is publicly available at https://github.com/zhangzhongbo2213/Attention-DP3.

**Keywords:** Embodied AI · 3D Point Cloud · Diffusion Policy

## 1 Introduction

Recent progress in imitation learning (IL) [2–5, 17, 43, 48] has enabled robot policies to solve complex manipulation tasks with improved stability and generalization. In particular, diffusion-based policies [5,19,25,37] generate actions via iterative denoising, providing an expressive and robust framework for multimodal behavior synthesis. Recent 3D diffusion variants [15,16,18,22,45] further improve generalization by conditioning on point clouds, leveraging explicit geometry to better handle viewpoint changes and out-of-distribution object configurations.

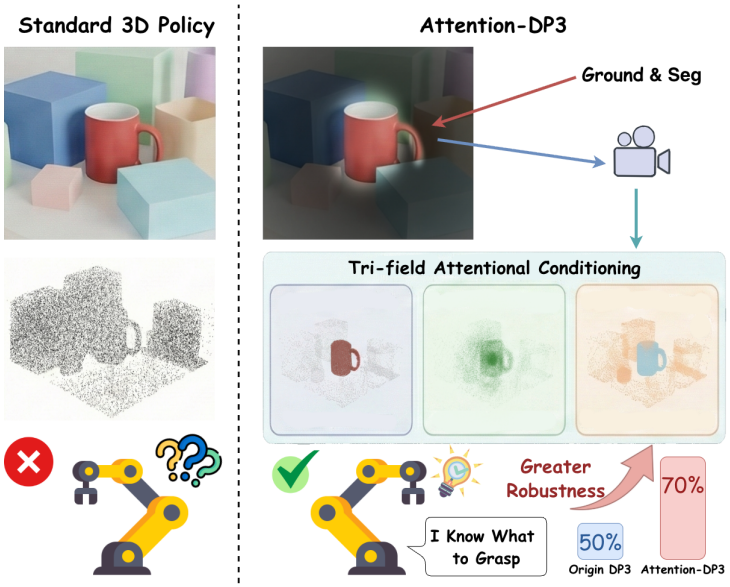



**Fig. 1: Standard 3D policies vs. Attention-DP3. (Left)** In cluttered scenes, sparse point clouds can be perceptually ambiguous under occlusion and distractors, causing unreliable target association. **(Right)** Attention-DP3 lifts object masks from RGB into 3D and uses geometry-aligned, object-aware attentional conditioning to guide action generation, helping the policy focus on what to grasp and where it is in 3D.

Despite these advances, manipulation in unstructured environments remains bottlenecked by perception [45]. While point clouds provide metric spatial constraints, they are sparse and highly ambiguous under clutter: when targets are partially occluded or tightly intermingled with distractors, the observed 3D points no longer cleanly separate the object of interest from background. This ambiguity induces a recurring failure mode for 3D diffusion policies, where the policy cannot reliably bind the language-specified target to its 3D extent, leading to unstable grasps or actions drifting toward distractors as clutter increases.

A natural direction is to exploit object-level cues from RGB observations to disambiguate the 3D scene [10, 13, 26, 34, 44]. However, directly fusing dense RGB features with sparse point-cloud representations can be brittle in clutter and may compromise geometric precision. Instead, we treat object-level cues as lightweight attentional prompts that modulate point-cloud-conditioned policy learning while preserving the underlying 3D diffusion backbone.

We propose **Attention-DP3**, a new 3D diffusion policy model that resolves clutter-induced perceptual ambiguity via object-aware 3D attention, while keeping the DP3 diffusion formulation unchanged. Given a language goal, our pipeline first performs open-vocabulary 2D segmentation on RGB images, then lifts the predicted target masks into 3D using calibrated camera geometry to obtain geometry-aligned object priors. We introduce Tri-Field Attentional Condition-

ing, which decomposes object-aware guidance into three complementary fields over 3D points: (i) a targetness field that anchors target points, (ii) an intra-target saliency field that emphasizes task-relevant geometry within the target, and (iii) a backgroundness field that suppresses distractors while retaining sufficient context for occlusion reasoning. These fields act as soft attention cues, enabling object-aware disambiguation without altering the diffusion backbone.

We evaluate Attention-DP3 on Adroit [31], DexArt [1], MetaWorld [42], and the real-world SO101 platform, where it consistently improves over DP3. Beyond average gains, we conduct stress tests that systematically scale visual clutter by increasing distractor objects at test time. Notably, DP3 exhibits a sharp performance drop as clutter intensifies, whereas Attention-DP3 remains stable and achieves gains of up to +31% under extreme distraction, highlighting strong zero-shot robustness to unseen clutter.

Our contributions are threefold. **(1) Spatially object-aware prompting for 3D diffusion policies:** we lift open-vocabulary 2D object masks into 3D and use them as geometry-aligned attentional prompts for point-cloud-conditioned policy learning, without modifying the diffusion backbone. **(2) Tri-Field Attentional Conditioning:** we propose a lightweight multi-field conditioning design decomposing object-aware guidance into target anchoring, intra-target structure emphasis, and distractor suppression for robust perception under clutter. **(3) Consistent gains and strong clutter robustness:** we demonstrate improvements across multiple benchmarks and substantial zero-shot robustness under systematically increased unseen clutter, including extreme distraction scenarios. The code is publicly available at https://github.com/zhangzhongbo2213/Attention-DP3.

## 2 Related Work

### 2.1 3D Representations for Robotic Manipulation

Point cloud learning has become a standard approach for 3D robotic perception. Early works such as PointNet [28] and PointNet++ [29] processed raw point clouds directly with permutation-invariant architectures. Subsequent methods incorporate local geometry with convolution-style operators (e.g., KPConv [38]) or attention mechanisms (e.g., Point Transformer [46]), improving feature expressiveness for 3D understanding. In manipulation settings, point-cloud encoders frequently serve as lightweight perception backbones within policies [30, 35, 45]. Compared to 2D-only representations, explicit 3D geometry reduces viewpoint sensitivity and provides accurate spatial constraints for contact-rich control [8, 9, 44]. However, pure geometry lacks semantic grounding, motivating its combination with task-relevant semantic cues.

### 2.2 Diffusion Policies for Imitation Learning

Diffusion policy models action generation as an iterative denoising process, providing a stable and expressive framework for learning multimodal behaviors from

demonstrations. Recent works extend diffusion policies to 3D observations by conditioning on point clouds, typically encoding them into compact latent vectors to guide a conditional denoiser [15,16,18,22,45]. This family of approaches has shown strong generalization across manipulation tasks and has inspired further improvements in robustness and scaling, including equivariant architectures (e.g., EquiBot [40]) and large-scale pre-training (e.g., FP3 [41]). Our method is built on this line of 3D diffusion policies, but augments the conditioning with a semantic prior that is explicitly aligned to the observed geometry, rather than learned through joint feature fusion.

### 2.3 Semantic-Guided 3D Control

Recent works increasingly use 2D foundation models to provide strong semantic priors for robotic manipulation [20, 32]. However, integrating these 2D signals into 3D control or attention mechanisms typically relies on learned cross-modal feature fusion [14] or explicit mask-based cropping [12,47]. These approaches can be brittle under visual clutter and often compromise strict spatial precision required for delicate tasks. Unlike methods requiring complex learned alignments or heavy pre-processing, we bypass 2D-3D feature fusion. We instead apply explicit geometric lifting to construct a training-free, geometry-aligned semantic attention signal, providing robust and lightweight guidance for 3D policies.

## 3 Method

### 3.1 Overview

As shown in Fig. 2, Attention-DP3 injects language-conditioned object awareness into DP3 through a geometry-aligned tri-field attention. At each timestep, the policy observes an RGB image $I$, a point cloud $P = \{p_i\}_{i=1}^{N}$, and proprioception $s$. A 3D perception encoder extracts a compact geometry feature $z_{\text{pc}}$, while a lightweight state encoder maps $s$ to $z_{\text{state}}$. In parallel, we design an Attn-Enhanced Module. A frozen grounding-and-segmentation pipeline (Grounding DINO + SAM2) [20,32] predicts a text-conditioned 2D mask on $I$, which we lift onto $P$ to obtain three geometry-aligned fields: Targetness, Intra-target Saliency, and Backgroundness. These fields are individually encoded by field encoders and subsequently aggregated into an attention feature $z_{\text{attn}}$ via an MLP. The diffusion denoiser (conditional U-Net) is conditioned on

$$c = [z_{\text{pc}},\, z_{\text{state}},\, z_{\text{attn}}], \tag{1}$$

and generates an $H$-step action trajectory via iterative denoising. By grounding language cues in 3D fields instead of learning implicit RGB–3D alignment, the policy remains robust in cluttered scenes. The tri-field design factorizes complementary roles: binding to the target (Targetness), emphasizing within-target structure (Intra-target Saliency), and suppressing distractors while retaining context (Backgroundness).

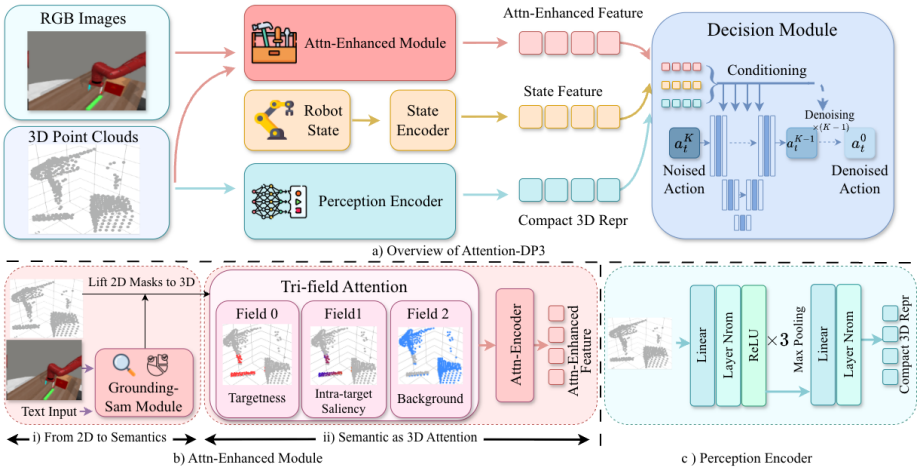



**Fig. 2: Attention-DP3 overview. (a) Pipeline.** The policy conditions on synchronized RGB, 3D point clouds, and proprioception. A 3D Perception Module encodes global geometry, a State Encoder embeds proprioception, and an Attn-Enhanced Module maps open-vocabulary 2D cues to an attention feature. Their embeddings condition a diffusion Decision Module to denoise an action trajectory. **(b) Attn-Enhanced Module.** A frozen grounding-and-segmentation pipeline predicts a text-conditioned 2D mask, which is lifted to three geometry-aligned fields: Targetness, Intra-target Saliency, and Backgroundness. These fields are individually encoded by Field Encoders and subsequently aggregated via an MLP. **(c) Perception Encoder.** A lightweight point-cloud backbone encodes raw point clouds into a compact 3D representation.

### 3.2 Explicit 2D-to-3D Object-Cue Lifting

A central challenge in cluttered manipulation is that sparse point clouds alone often do not cleanly separate the object of interest from surrounding distractors, particularly under occlusion. A common approach is to learn implicit RGB–3D alignment in a shared feature space and fuse representations across modalities. However, this learned alignment can be brittle in heavy clutter and can blur geometry that is critical for precise control. We adopt an explicit lifting strategy instead: a text-conditioned object mask is extracted in 2D and deterministically mapped onto the observed 3D points. Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and a language query, a frozen segmentation model outputs a binary mask $M \in \{0,1\}^{H\times W}$ that marks pixels belonging to the queried object. With calibrated camera intrinsics and extrinsics, each 3D point $p_i$ is projected onto the image plane:

$$(u_i, v_i) = \pi(p_i), \tag{2}$$

Here $\pi(\cdot)$ is the calibrated 3D-to-2D projection operator induced by the camera intrinsics and extrinsics: it maps a 3D point to its corresponding pixel coordinate under the pinhole camera model (with points behind the camera or outside the image treated as background).

For mask lookup, we discretize the continuous projection to the nearest pixel index. Each point then receives a lifted object indicator through nearest-neighbor lookup:

$$m_i = M(u_i, v_i), \qquad m_i \in \{0, 1\}. \tag{3}$$

The resulting $\{m_i\}_{i=1}^{N}$ provides a geometry-aligned object cue over the point set: it is derived from RGB while remaining strictly consistent with the observed 3D geometry via calibrated projection. This explicit lifting avoids learning 2D–3D correspondence in an embedding space and offers a robust bridge from RGB object cues to point-cloud policy conditioning. We then refine the lifted cue into a tri-field attention representation to better handle clutter and occlusion.

### 3.3 Geometry-aligned Lifted Tri-Field Attention

From the lifted indicator $\{m_i\}$, we construct a deterministic Lifted Tri-Field Attention (LTFA) signal $A \in \mathbb{R}^{3\times N}$. LTFA appends three complementary per-point channels to the point cloud while leaving the 3D coordinates unchanged, thereby injecting object-aware attention without altering any geometric measurements. For each point $p_i$, we define $A_{\cdot,i} = [A_{\mathrm{T},i}, A_{\mathrm{S},i}, A_{\mathrm{B},i}]^\top$, where each component is a scalar field designed to stabilize conditioning under clutter and occlusion.

**Field 1: Targetness.** We define a targetness field that provides an explicit where-to-attend signal in 3D:

$$A_{\mathrm{T},i} = m_i. \tag{4}$$

**Field 2: Intra-target saliency.** Binary masks can be noisy or fragmented and may miss thin parts under occlusion. To provide a graded structural cue within the target region, we compute a normalized distance field $D(u, v) \in [0, 1]$ from the 2D mask geometry and lift it to points:

$$s_i = D(u_i, v_i), \qquad A_{\mathrm{S},i} = m_i \cdot s_i. \tag{5}$$

In practice, $D$ is obtained via the Euclidean distance transform inside the mask and normalized to $[0, 1]$. This field softly emphasizes geometrically central or contact-relevant subregions instead of uniformly weighting all target points.

**Field 3: Backgroundness.** Rather than discarding non-target points, we encode complementary context via a backgroundness field:

$$A_{\mathrm{B},i} = 1 - m_i. \tag{6}$$

This encourages the downstream encoder to down-weight distractor evidence while preserving enough context for reasoning about occlusion and scene layout.

**Field encoder.** While LTFA itself is deterministic and training-free, we learn a lightweight field encoder $E_{\text{field}}$ (point-wise MLP + symmetric pooling) to aggregate per-point fields into a compact vector:

$$z_{\text{attn}} = E_{\text{field}}(A). \tag{7}$$

Treating $A$ as a three-channel per-point feature, we apply a shared MLP to each column $A_{\cdot,i}$ and pool across points, yielding a global attention feature. Grounding/segmentation models, mask lifting, and LTFA construction (Eqs. (2) to (6)) remain frozen; only $E_{\text{field}}$ and downstream DP3 components are learned.

### 3.4 Attention-conditioned Diffusion Policy

We follow the DP3 formulation and condition the denoiser on $c$ (Eq. (1)). Let $\mathbf{x}_0 \in \mathbb{R}^{H \times d_a}$ denote the clean action sequence (horizon $H$, action dimension $d_a$). The forward diffusion process corrupts $\mathbf{x}_0$ with Gaussian noise, producing $\mathbf{x}_t$ at diffusion step $t \in \{1, \ldots, T\}$:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\mathbf{x}_0,\ (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{8}$$

where $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$ is determined by a fixed noise schedule $\{\alpha_t\}_{t=1}^{T}$.

The policy is parameterized by a conditional denoiser $\epsilon_\theta$. Following DP3, we instantiate $\epsilon_\theta$ as a conditional U-Net over the temporally structured action sequence, where the diffusion step $t$ is embedded and injected into each residual block. The global condition $c$ is incorporated via FiLM-style feature modulation in intermediate layers, enabling the denoiser to couple action generation with both geometry features and the proposed tri-field attention. Formally, the denoiser predicts the injected noise:

$$\hat{\epsilon} = \epsilon_\theta(\mathbf{x}_t,\ t,\ c). \tag{9}$$

We train with the standard noise-prediction objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0, t, \epsilon}\Big[\|\epsilon - \epsilon_\theta(\mathbf{x}_t,\ t,\ c)\|_2^2\Big]\,, \tag{10}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{x}_t$ is sampled via Eq. (8). At inference, we start from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively apply $\epsilon_\theta$ from $t = T$ to 1, yielding the final action trajectory $\mathbf{x}_0$.

## 4 Experiments

### 4.1 Experiment Setup

**Benchmarks.** Adroit [31] contains highly challenging 24-DoF anthropomorphic hand manipulation tasks. DexArt [1] focuses on articulated-object manipulation, which requires precise spatial reasoning. MetaWorld [42] provides diverse and structurally varied tabletop manipulation tasks, such as Pick-Place, Box-Close, and Shelf-Place. SO101 includes Place Cube, Push Cube, and Stack Cube.

**Data Collection.** We obtain expert demonstrations using a scripted policy for MetaWorld, and employ VRL3 [39] and PPO [33] for the Adroit and DexArt environments, respectively. Our dataset includes 10 episodes per task for the Adroit and MetaWorld benchmarks, and 100 episodes for DexArt. For SO101, we collect 10 expert trajectories via LeRobot teleoperation using a global camera at 30 Hz; each SO101 arm has 6 degrees of freedom.

**Baselines.** We evaluate Attention-DP3 against a wide spectrum of state-of-the-art visuomotor policies, which can be categorized into three groups: (i) Implicit Behavioral Cloning Baselines: including BCRNN [24] and Implicit Behavioral Cloning (IBC) [6]; (ii) Image-based Generative Policies: including the foundational Diffusion Policy (DP) [5], Consistency Policy (CP) [27], AdaFlow [11], and the recently proposed Vision-to-Action flow matching policy (VITA) [7]; (iii) 3D Point Cloud-based Policies: including 3D Diffusion Policy (DP3) [45] and its streamlined variant (Simple DP3), as well as contemporary SOTA spatial-aware methods such as $H^3$DP [23] and FreqPolicy [36].

**Observations and text queries.** Each policy receives synchronized RGB images, depth-reconstructed point clouds (with calibrated camera intrinsics/extrinsics), and proprioception. For Attention-DP3, the text query is a fixed task-level description of the manipulation target (e.g., the object name). We use the same query for training and evaluation, avoiding test-time prompt tuning.

**Training and evaluation protocol.** Following standard practice in sample-efficient imitation learning, we train all methods with 10 expert demonstrations per task, except on DexArt, where we use 100 demonstrations because of the complexity of articulated-object manipulation. We optimize all models with AdamW [21] using a learning rate of $10^{-4}$ ($\beta = (0.95, 0.999)$, $\epsilon = 10^{-8}$, weight decay $10^{-6}$), a batch size of 128, and no gradient accumulation. We train for 1000 epochs on MetaWorld and for 3000 epochs on the other benchmarks. We assess 20 episodes every 200 epochs and calculate the mean of the top five success rates. To evaluate clutter generalization, we introduce distractor objects with progressively increasing severity to create cluttered conditions unseen during training, while keeping the task goal unchanged (Sec. 4.4). All implementations are built in PyTorch and evaluated on a single NVIDIA RTX 3090 GPU.

### 4.2 Results on Simulation Benchmarks

Table 1 and Table 2 present the quantitative comparison of Attention-DP3 against strong baselines. In the highly challenging MetaWorld suite, our Attention-DP3 achieves an average success rate of 0.73, which is higher than the score of DP3 and VITA. On spatial-reasoning-intensive tasks such as *Push-Wall* and *Pick-Place*, Attention-DP3 provides large gains of +0.43 and +0.42 over DP3,

**Table 1: Success Rates on Adroit and DexArt Benchmarks.** Missing entries (–) indicate methods not evaluated on that benchmark.

| Method | Adroit | | | Avg | DexArt | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|
| | Door | Pen | Hammer | | Laptop | Faucet | Toilet | Bucket | |
| IBC [6] | 0.00 | 0.00 | 0.09 | 0.03 | 0.03 | 0.07 | 0.00 | 0.14 | 0.06 |
| BCRNN [24] | 0.00 | 0.00 | 0.09 | 0.03 | 0.03 | 0.01 | 0.00 | 0.05 | 0.02 |
| AdaFlow [11] | 0.27 | 0.18 | 0.45 | 0.30 | – | – | – | – | – |
| CP [27] | 0.31 | 0.13 | 0.45 | 0.30 | – | – | – | – | – |
| $H^3$DP [23] | – | – | – | – | 0.81 | 0.34 | 0.70 | **0.28** | 0.53 |
| FreqPolicy [36] | – | – | – | – | **0.85** | 0.30 | **0.77** | 0.25 | 0.54 |
| DP [5] | 0.37 | 0.13 | 0.45 | 0.32 | 0.69 | 0.23 | 0.58 | 0.20 | 0.43 |
| DP3 [45] | 0.62 | 0.43 | **1.00** | 0.68 | 0.77 | 0.36 | 0.70 | 0.25 | 0.52 |
| Simple DP3 [45] | 0.58 | 0.46 | **1.00** | 0.68 | 0.79 | 0.26 | 0.63 | 0.22 | 0.48 |
| VITA [7] | 0.81 | **0.55** | 0.96 | 0.77 | 0.82 | **0.39** | 0.72 | **0.28** | 0.55 |
| **Ours** | **0.83** | 0.50 | **1.00** | **0.78** | **0.85** | 0.37 | 0.72 | **0.28** | **0.56** |

**Table 2: MetaWorld results by difficulty with representative tasks.** For each difficulty, we show two representative tasks plus an ellipsis indicating additional tasks in the same split, and report the difficulty-average (Avg) over all tasks in that split; Overall is the mean over all MetaWorld tasks. Full per-task results are provided in the supplementary material.

| Method | Easy | | | | Medium | | | | Hard | | | | Very Hard | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | handle-pull | handle-pull-side | ⋯ | Avg | push-wall | hammer | ⋯ | Avg | pick-place | push | ⋯ | Avg | pick-place-wall | disassemble | ⋯ | Avg | |
| DP | 0.27 | 0.23 | – | 0.836 | 0.20 | 0.15 | – | 0.311 | 0.00 | 0.30 | – | 0.090 | 0.05 | 0.43 | – | 0.266 | 0.574 |
| DP3 | 0.34 | 0.53 | – | 0.857 | 0.49 | 0.76 | – | 0.525 | 0.12 | 0.51 | – | 0.248 | 0.35 | 0.69 | – | 0.442 | 0.669 |
| VITA | 0.20 | 0.52 | – | 0.839 | 0.49 | 0.90 | – | 0.546 | 0.15 | 0.70 | – | 0.315 | 0.48 | 0.61 | – | 0.548 | 0.683 |
| Ours | **0.46** | **0.62** | – | **0.870** | **0.92** | **0.95** | – | **0.599** | **0.54** | **0.82** | – | **0.378** | **0.60** | **0.86** | – | **0.620** | **0.726** |

respectively, which indicates that the proposed semantic prior guides diffusion toward task-relevant geometric features. We observe similarly consistent improvements on Adroit (0.78 average versus 0.68 for DP3) and DexArt (0.56 average versus 0.52 for DP3), confirming that the benefit of semantic attention extends to both dexterous manipulation and articulated-object manipulation.

### 4.3 Real-world Experiment Results

We further validate Attention-DP3 on real-robot SO101 platform. Figure 3 shows the real-world experiment setup. We evaluate three representative tabletop tasks. In *Place Cube*, the robot places the cube into the bowl. In *Push Cube*, success is achieved when the blue cube is pushed to within 2 cm of the red cube. In *Stack Cube*, success is achieved when the blue cube is placed stably on top of the red cube. We report success rates over repeated 20 rollouts under the same evaluation protocol for both methods.

As shown in Table 3, Attention-DP3 improves all three tasks with the largest gain on *Push Cube*, which requires precise spatial reasoning under local contact

**Table 3: Real-world results on SO101.** Attention-DP3 consistently outperforms DP3 across all tasks.

| **Task** | Stack Cube | Push Cube | Place Cube | **Average** |
|---|---|---|---|---|
| **DP3** | 0.50 | 0.30 | 0.75 | 0.52 |
| **Attention-DP3** | **0.70** | **0.65** | **0.85** | **0.73** |

changes. The average success rate improves from 0.52 to 0.73, which supports that geometry-aligned attentional conditioning transfers effectively from simulation to real-world manipulation.

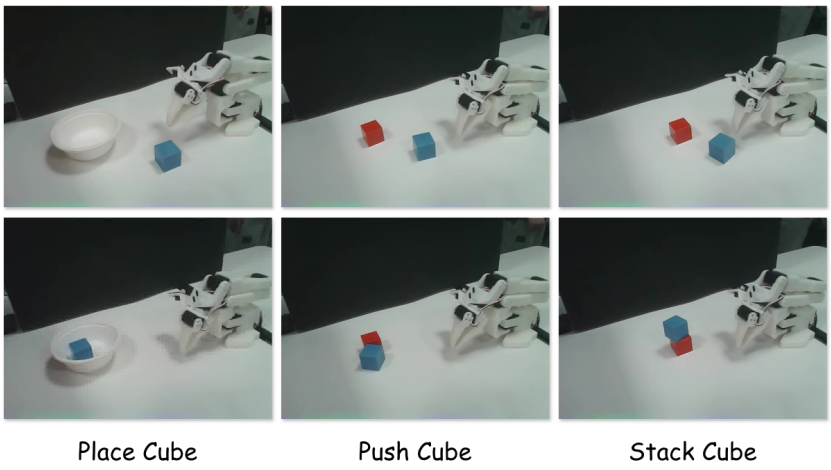
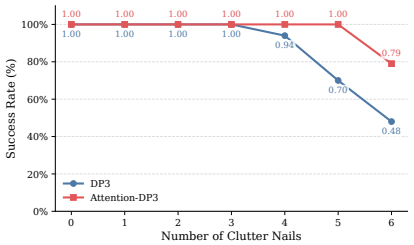



**Fig. 3: Real-world experimental setting on SO101.** We evaluate three tasks, including Place Cube, Push Cube, and Stack Cube.

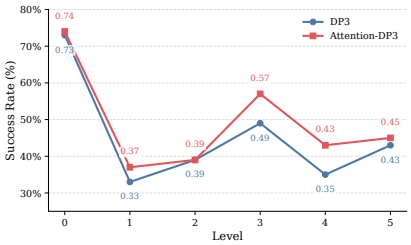
### 4.4 Robustness to Visual Clutter

We evaluate visual-clutter robustness in both simulation and real-world settings.

To assess the robustness of the policies in cluttered environments, we design generalization scenarios characterized by progressively increasing levels of distraction. For the Adroit *Hammer* task, we vary the number of distractor nails from 0 to 6. Similarly, for the MetaWorld *Stick-Push* task, we define six distinct distraction levels based on the quantity of distractor blocks, ranging from 0 to 5. The 0-block setting represents a clean, distractor-free environment, while the subsequent levels correspond to the introduction of 1 to 5 unseen blocks randomly scattered across the functional workspace of the robot. Detailed configurations of these experimental setups are provided in the supplementary material.




**(a) Adroit Hammer (Clutter: Nails).** Success rate under increasing distractor nails.




**(b) MetaWorld Stick-Push (Clutter: Blocks).** Success rate under increasing distractor blocks.

**Fig. 4: Robustness to visual clutter.** Our attention mechanism improves robustness under unseen distractors across both tasks.

**Table 4: Real-world clutter robustness on SO101.** Average success rate across Place Cube, Push Cube, and Stack Cube under progressively added distractors.

| Method | Orig. | +Clip | +Stick | +Extra |
|---|---|---|---|---|
| DP3 | 0.52 | 0.43 | 0.38 | 0.13 |
| Attention-DP3 | **0.73** | **0.72** | **0.65** | **0.45** |

As shown in Figure 4, on *Adroit Hammer*, both methods are nearly unaffected by light clutter. However, their behaviors diverge significantly in the high-density regime (4–6 nails). DP3 exhibits a clear monotonic collapse, indicating its susceptibility to task-irrelevant geometry. In contrast, Attention-DP3 preserves a remarkably flat degradation curve, maintaining saturated success through moderate clutter and demonstrating substantially higher tolerance before failure.

A similar pattern is observed in the *MetaWorld Stick-Push* task across varying density levels (0–5 blocks). As the number of blocks increases, the performance of DP3 exhibits fluctuations and a general downward trend, indicating unstable foreground extraction under distribution shifts. In contrast, Attention-DP3 consistently outperforms DP3 across almost all clutter levels. The most pronounced performance gains are observed at mid-to-high clutter levels (3–5 blocks), where distractors severely interfere with local geometric perception.

We further perform a real-world clutter stress test by progressively adding distractors to the SO101 scenes, including clips, sticks, and task-specific extra distractors. Table 4 reports the average success rate across the three real-world tasks. While DP3 degrades substantially as clutter increases, Attention-DP3 maintains a higher success rate under all clutter settings, indicating that object-aware conditioning helps the policy preserve target focus in unstructured real-world scenes.

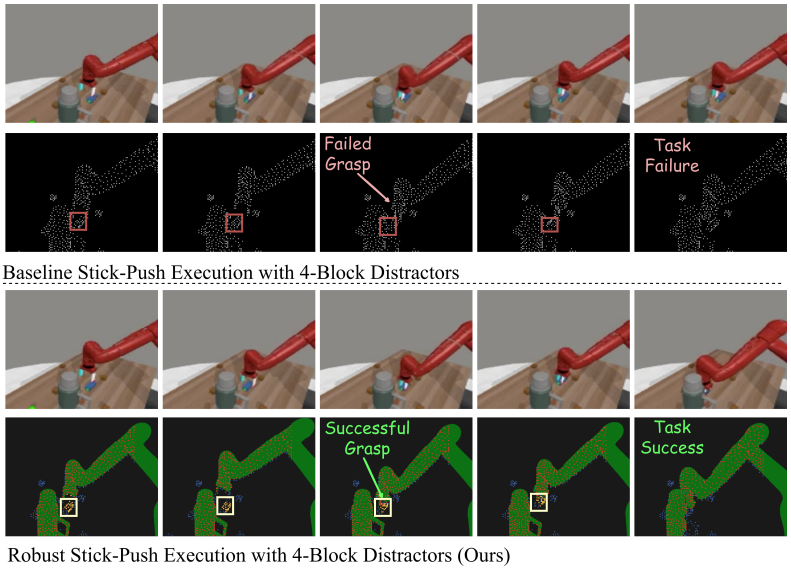



**Fig. 5: Qualitative comparison on MetaWorld Stick-Push with four distractor blocks.** Top: DP3; bottom: Attention-DP3.

**Table 5: Ablation on semantic attention fields.** We report success rates on Adroit *Door* and MetaWorld tasks. **Avg.** is the mean over the four tasks.

| Variant | T | I | B | Door | Stick-Pull | Pick-Place | Push-Wall | Avg. |
|---|---|---|---|---|---|---|---|---|
| Full | ✓ | ✓ | ✓ | **0.83** | **0.59** | **0.54** | **0.92** | **0.72** |
| T only | ✓ | | | 0.74 | 0.53 | 0.33 | 0.57 | 0.5425 |
| I only | | ✓ | | 0.73 | 0.51 | 0.32 | 0.64 | 0.55 |
| B only | | | ✓ | 0.68 | 0.50 | 0.33 | 0.63 | 0.535 |
| T+I | ✓ | ✓ | | 0.80 | 0.30 | 0.39 | 0.58 | 0.5175 |
| T+B | ✓ | | ✓ | 0.75 | 0.56 | 0.36 | 0.58 | 0.5625 |
| B+I | | ✓ | ✓ | 0.82 | 0.54 | 0.38 | 0.64 | 0.595 |

## 5 Ablation Study

### 5.1 Ablation on attention fields.

Table 5 compares different field combinations: Targetness (T), Intra-target saliency (I), and Backgroundness (B). Single-field variants underperform, two-field variants show task-dependent trade-offs, and the complete three-field configuration yields the most consistent improvements and highest average, validating the full design.

### 5.2 Ablation on fusion strategies.

Table 6 compares early fusion (concatenating attention and geometry at input) with late fusion. Late fusion performs best, especially when each field is encoded

by an independent AttnEncoder before fusion with geometric features, indicating that field-wise encoding preserves complementary cues and reduces cross-field interference. Since all fusion variants use the same lifted mask source, the gap is not explained by external semantics alone; separate geometry-aligned fields and late fusion are important for turning the semantic prior into effective 3D control.

**Table 6: Ablation on fusion strategy.** We report success rates on Adroit *Door* and MetaWorld tasks. **Avg.** is the mean over the four tasks.

| Fusion Strategy | Encoder | Door | Stick-Pull | Pick-Place | Push-Wall | Avg. |
|---|---|---|---|---|---|---|
| Late fusion | 3 AttnEncoders + DP3Encoder | **0.83** | **0.59** | **0.54** | **0.92** | **0.72** |
| Late fusion | Single AttnEncoder + DP3Encoder | 0.69 | 0.42 | 0.36 | 0.62 | 0.5225 |
| Early fusion | Attn+PC concat → DP3 Encoder | 0.66 | 0.02 | 0.07 | 0.27 | 0.255 |

### 5.3 Inference and Operational Latency

We analyze the operational bandwidth of Attention-DP3 via control latency. To isolate the overhead from semantic attention branches under asynchronous execution, we define the *effective control step latency* $L_{step}$ as the average time per executed action step. With prediction horizon $H = 16$ and execution window $k = 8$, each inference cycle spans $T_{phys} = 0.8\,\text{s}$ of physical execution at 10 Hz. We compute

$$L_{step} = \frac{T_{phys} + T_{inf}}{k}, \tag{11}$$

where $T_{inf}$ is the end-to-end inference time, including the Grounded-SAM-2 (GS2) perception pipeline.

Table 7 shows that Attention-DP3 preserves smooth control bandwidth. Its mean $L_{step}$ is 0.198 s, versus 0.122 s for DP3, a modest increase without visible stuttering in asynchronous deployment. The policy-side overhead is minimal: compared with DP3, Attention-DP3 adds only 0.001 s per step, 0.01 GB peak memory, and 3.8M parameters. The remaining end-to-end cost comes from the frozen GS2 perception branch, whose lighter Swin-T variant reduces perception latency from 0.100 s to 0.044 s.

### 5.4 Diagnostic Analysis of Failure Modes

We evaluate whether failures of Attention-DP3 are primarily caused by 2D grounding errors. We consider three tasks: Coffee-Push, Peg-Unplug-Side, and Reach-Wall. For each task, we select 10 failure episodes of Attention-DP3 and rerun DP3 using identical random seeds. Table 8 shows that most Attention-DP3 failures in Coffee-Push and Peg-Unplug-Side come from missed detections by Grounding DINO (8/10 and 7/10), while DP3 also fails frequently on the same episodes (7/8 and 5/7), indicating inherently difficult cases rather than

**Table 7: Effective Control Step Latency.** Per-step latency $L_{step}$ for execution window $k = 8$ ($T_{phys} = 0.8\,\text{s}$), averaging physical execution and inference latency $T_{inf}$. Measurements use one NVIDIA RTX 3090 GPU.

| Task | DP3 (Baseline) | Attn-DP3 (Ours) | Latency Gap |
|---|---|---|---|
| Adroit Pen | 0.120 s | 0.213 s | +0.093 s |
| Adroit Door | 0.124 s | 0.174 s | +0.050 s |
| Adroit Hammer | 0.123 s | 0.209 s | +0.086 s |
| **Mean** | **0.122 s** | **0.198 s** | **+0.076 s** |

Attention-DP3-specific failures. In Reach-Wall, grounding is always correct (0/10), yet DP3 fails in all episodes (10/10), suggesting physical feasibility as the bottleneck. Thus, grounding errors explain many cluttered-task failures, but correct grounding does not eliminate physical constraints.

**Table 8: Diagnostic Analysis of Failure Mode Overlap.** We isolate 10 Attention-DP3 failure episodes and rerun DP3 with identical seeds. "Overlap Fail" reports DP3 failures when Attention-DP3 suffers Grounding DINO errors.

| Task | Total Failures | Attn-DP3 (Fail via DINO) | DP3 (Overlap Fail) | DP3 (Total Fail) |
|---|---|---|---|---|
| *Coffee-Push* | 10 | 8 | 7 / 8 | 9 / 10 |
| *Peg-Unplug-Side* | 10 | 7 | 5 / 7 | 8 / 10 |
| *Reach-Wall* | 10 | 0 | N/A | 10 / 10 |

### 5.5 Robustness to Imperfect 2D Perception

We further examine whether Attention-DP3 depends critically on accurate 2D segmentation. For each of five tasks, we evaluate 100 rollouts with matched initial seeds for Attention-DP3 and DP3, and categorize each episode by Attention-DP3 key-step segmentation quality. With incorrect segmentation, Attention-DP3 remains comparable to DP3 (29/134 vs. 30/134 successful rollouts), suggesting that erroneous masks do not introduce disproportionate failures. With correct segmentation, Attention-DP3 substantially outperforms DP3 (233/366 vs. 155/366), indicating that reliable object cues are effectively exploited. Direct mask perturbation shows the same trend: replacing predicted masks with ground-truth masks only improves average success from 0.524 to 0.542, while random mask dropout degrades smoothly and remains above DP3 even at $p = 0.9$ (0.456 vs. 0.370). We also test prompt sensitivity with five generated prompts per task; generated and manual prompts perform comparably (0.73 vs. 0.75 mIoU, 0.55 vs. 0.56 success rate), suggesting that prompts mainly specify the target object rather than drive policy gains.

### 5.6 Robustness to Perceptual Foundation Models

To test robustness to the perception backbone, we vary the grounding model used by Attention-DP3. Concretely, we replace the default Grounding DINO Swin-B with a lightweight Swin-T variant, denoted Attention-DP3 (Swin-T), and compare it with the standard Attention-DP3 and the DP3 baseline. As shown in Table 9, downsizing the grounding model yields comparable success rates despite reduced detection precision. Notably, Attention-DP3 (Swin-T) reaches 100% success on Hammer, outperforming both the standard model and DP3. On spatial reasoning tasks such as Push-Wall and Pick-Place, it remains markedly stronger than the geometry-only baseline. Overall, these results suggest that our semantic-aware design is resilient to perceptual noise and does not rely on large or expensive vision backbones to achieve strong manipulation performance.

**Table 9: Robustness to Perceptual Foundation Models.** Success rates of DP3, Attention-DP3, and Attention-DP3 (Swin-T) using a lightweight Grounding DINO Swin-T backbone.

| Method | Hammer | Box-Close | Push-Wall | Pick-Place | Shelf-Place | Stick-Pull |
|---|---|---|---|---|---|---|
| DP3 | 0.76 | 0.42 | 0.49 | 0.12 | 0.17 | 0.27 |
| Attn-DP3 (Swin-B) | 0.95 | 0.52 | 0.92 | 0.54 | 0.22 | 0.59 |
| Attn-DP3 (Swin-T) | **1.00** | 0.43 | 0.84 | 0.47 | 0.17 | 0.53 |

## 6 Conclusion

We presented Attention-DP3, a spatially object-aware 3D diffusion policy that mitigates clutter-induced perceptual ambiguity without altering the DP3 diffusion formulation. By lifting open-vocabulary 2D masks into geometry-aligned 3D object priors and applying Tri-Field Attentional Conditioning, it reliably binds language-specified targets to their 3D extents and remains robust under occlusion and heavy distraction. Experiments on Adroit, DexArt, MetaWorld, and the real-world SO101 platform show consistent gains over DP3, and clutter stress tests confirm strong zero-shot stability as distractors scale. We hope this work highlights lightweight, geometry-preserving object-level prompting as a practical route to robust manipulation in unstructured environments.

## 7 Acknowledgment

This paper is supported by the National Natural Science Foundation of China (62422610, U23A20386, 62441231, 62276045, 62576072), Liao Ning Science and Technology Plan (2025JH2/101330121, 2025JH2/101330124, 2023JH26/10200016), and Dalian City Science and Technology Innovation Fund (2023JJ11CG001).

## References


1. Bao, C., Xu, H., Qin, Y., Wang, X.: Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21190–21200 (2023) 3, 7
2. Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M.R., Finn, C., Fusai, N., Galliker, M.Y., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. In: 9th Annual Conference on Robot Learning (2025) 1
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: $\pi_0$: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024) 1
4. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022) 1
5. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025) 1, 8, 9
6. Florence, P., Lynch, C., Zeng, A., Ramirez, O.A., Wahid, A., Downs, L., Wong, A., Lee, J., Mordatch, I., Tompson, J.: Implicit behavioral cloning. In: Conference on robot learning. pp. 158–168. PMLR (2022) 8, 9
7. Gao, D., Zhao, B., Lee, A., Chuang, I., Zhou, H., Wang, H., Zhao, Z., Zhang, J., Soltani, I.: Vita: Vision-to-action flow matching policy. arXiv preprint arXiv:2507.13231 (2025) 8, 9
8. Gervet, T., Xian, Z., Gkanatsios, N., Fragkiadaki, K.: Act3d: 3d feature field transformers for multi-task robotic manipulation. arXiv preprint arXiv:2306.17817 (2023) 3
9. Goyal, A., Xu, J., Guo, Y., Blukis, V., Chao, Y.W., Fox, D.: Rvt: Robotic view transformer for 3d object manipulation. In: Conference on Robot Learning. pp. 694–710. PMLR (2023) 3
10. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024) 2
11. Hu, X., Liu, Q., Liu, X., Liu, B.: Adaflow: Imitation learning with variance-adaptive flow-based policies. Advances in Neural Information Processing Systems **37**, 138836–138858 (2024) 8, 9
12. Huang, H., Chen, X., Chen, Y., Li, H., Han, X., Wang, Z., Wang, T., Pang, J., Zhao, Z.: Roboground: Robotic manipulation with grounded vision-language priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22540–22550 (2025) 4
13. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973 (2023) 2
14. Jia, Y., Liu, J., Chen, S., Gu, C., Wang, Z., Luo, L., Lee, L., Wang, P., Wang, Z., Zhang, R., et al.: Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. arXiv preprint arXiv:2411.18623 (2024) 4

15. Ke, T.W., Gkanatsios, N., Fragkiadaki, K.: 3d diffuser actor: Policy diffusion with 3d scene representations. arXiv preprint arXiv:2402.10885 (2024) 1, 4
16. Ke, T.W., Gkanatsios, N., Xu, J., Fragkiadaki, K.: Bi3d diffuser actor: 3d policy diffusion for bi-manual robot manipulation. In: CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data (2024) 1, 4
17. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024) 1
18. Li, H., Feng, Q., Zheng, Z., Feng, J., Chen, Z., Knoll, A.: Language-guided object-centric diffusion policy for generalizable and collision-aware manipulation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 12834–12841. IEEE (2025) 1, 4
19. Liang, Z., Mu, Y., Ma, H., Tomizuka, M., Ding, M., Luo, P.: Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16467–16476 (2024) 1
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 4
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 8
22. Lu, G., Gao, Z., Chen, T., Dai, W., Wang, Z., Ding, W., Tang, Y.: Manicm: Real-time 3d diffusion policy via consistency model for robotic manipulation. arXiv preprint arXiv:2406.01586 (2024) 1, 4
23. Lu, Y., Tian, Y., Yuan, Z., Wang, X., Hua, P., Xue, Z., Xu, H.: $H^3$DP: Triply-hierarchical diffusion policy for visuomotor learning. arXiv preprint arXiv:2505.07819 (2025) 8, 9
24. Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., Martín-Martín, R.: What matters in learning from offline human demonstrations for robot manipulation. arXiv preprint arXiv:2108.03298 (2021) 8, 9
25. Park, J., Kim, Y., Kim, S., Lee, B.J., Kim, S.: Diar: Diffusion-model-guided implicit q-learning with adaptive revaluation. arXiv preprint arXiv:2410.11338 (2024) 1
26. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 815–824 (2023) 2
27. Prasad, A., Lin, K., Wu, J., Zhou, L., Bohg, J.: Consistency policy: Accelerated visuomotor policies via consistency distillation. arXiv preprint arXiv:2405.07503 (2024) 8, 9
28. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 3
29. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017) 3
30. Qin, Y., Huang, B., Yin, Z.H., Su, H., Wang, X.: Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In: Conference on Robot Learning. pp. 594–605. PMLR (2023) 3

31. Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., Levine, S.: Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. arXiv preprint arXiv:1709.10087 (2017) 3, 7
32. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 4
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017) 8
34. Shen, W., Yang, G., Yu, A., Wong, J., Kaelbling, L.P., Isola, P.: Distilled feature fields enable few-shot language-guided manipulation. arXiv preprint arXiv:2308.07931 (2023) 2
35. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Conference on Robot Learning. pp. 785–799. PMLR (2023) 3
36. Su, Y., Liu, N., Chen, D., Zhao, Z., Wu, K., Li, M., Xu, Z., Che, Z., Tang, J.: Freqpolicy: Efficient flow-based visuomotor policy via frequency consistency. arXiv preprint arXiv:2506.08822 (2025) 8, 9
37. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024) 1
38. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019) 3
39. Wang, C., Luo, X., Ross, K., Li, D.: Vrl3: A data-driven framework for visual deep reinforcement learning. Advances in Neural Information Processing Systems **35**, 32974–32988 (2022) 8
40. Yang, J., Cao, Z.a., Deng, C., Antonova, R., Song, S., Bohg, J.: Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning. arXiv preprint arXiv:2407.01479 (2024) 4
41. Yang, R., Chen, G., Wen, C., Gao, Y.: Fp3: A 3d foundation policy for robotic manipulation. arXiv preprint arXiv:2503.08950 (2025) 4
42. Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., Levine, S.: Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Conference on robot learning. pp. 1094–1100. PMLR (2020) 3, 7
43. Zare, M., Kebria, P.M., Khosravi, A., Nahavandi, S.: A survey of imitation learning: Algorithms, recent developments, and challenges. IEEE Transactions on Cybernetics **54**(12), 7173–7186 (2024) 1
44. Ze, Y., Yan, G., Wu, Y.H., Macaluso, A., Ge, Y., Ye, J., Hansen, N., Li, L.E., Wang, X.: Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In: Conference on robot learning. pp. 284–301. PMLR (2023) 2, 3
45. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. arXiv preprint arXiv:2403.03954 (2024) 1, 2, 3, 4, 8, 9
46. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021) 3
47. Zhu, Y., et al.: Vima: General robot manipulation with multimodal prompts. In: International Conference on Learning Representations (ICLR) (2023) 4

48. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023) 1