GraspLDP: Towards Generalizable Grasping Policy via Latent Diffusion

Enda Xiang[1], Haoxiang Ma[1,2], Xinzhu Ma[1], Zicheng Liu[1,3], Di Huang[1†]

[1]School of Computer Science and Engineering, Beihang University
[2]Shanghai Artificial Intelligence Laboratory
[3]BCC Lab, Hangzhou International Innovation Institute, Beihang University

{endaxiang, mahaoxiang822, xinzhuma, liuzicheng, dhuang}@buaa.edu.cn

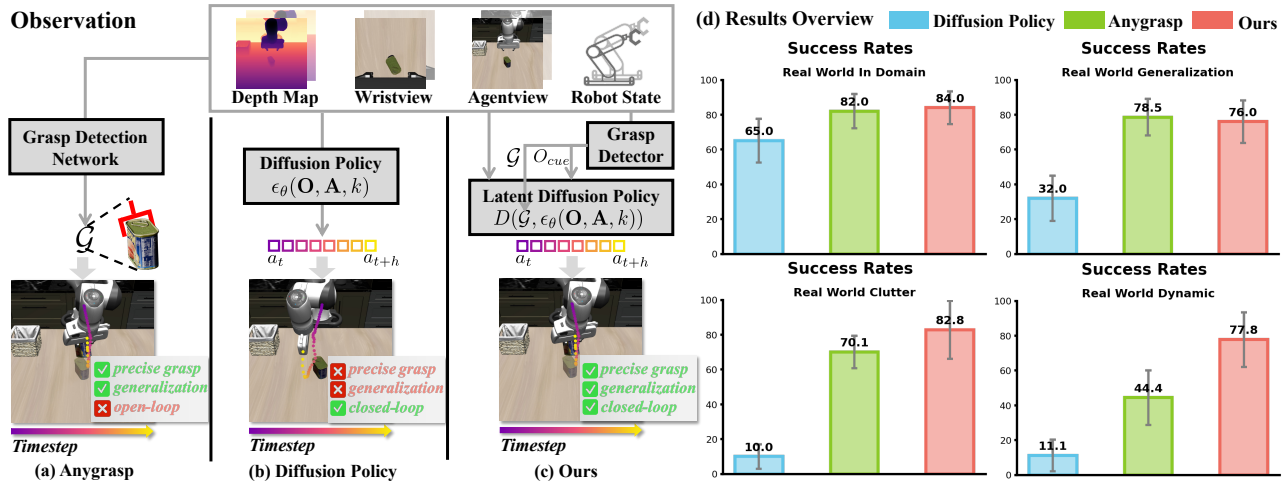

Figure 1. We introduce **GraspLDP**, a generalizable grasping policy integrated with the prior from grasp detector via latent diffusion. Specifically, prior works generally (*a*) predict the grasp pose (*e.g.* Anygrasp [10]) or (*b*) generate action sequence (*e.g.* Diffusion Policy [7]) for grasping. In contrast, (*c*) our method extracts grasp priors from a pre-trained grasp detector for action refinement in latent space, and (*d*) achieves substantial advantages over previous works in diverse grasping tasks.

Abstract

This paper focuses on enhancing the grasping precision and generalization of manipulation policies learned via imitation learning. Diffusion-based policy learning methods have recently become the mainstream approach for robotic manipulation tasks. As grasping is a critical subtask in manipulation, the ability of imitation-learned policies to execute precise and generalizable grasps merits particular attention. Existing imitation learning techniques for grasping often suffer from imprecise grasp executions, limited spatial generalization, and poor object generalization. To address these challenges, we incorporate grasp prior knowledge into the diffusion policy framework. In particular, we employ a latent diffusion policy to guide action chunk decoding with grasp pose prior, ensuring that generated motion trajectories adhere closely to feasible grasp configurations. Furthermore, we introduce a self-supervised reconstruction objective during diffusion to embed the graspness prior: at each reverse diffusion step, we reconstruct wrist-camera images back-projected the graspness from the intermediate representations. Both simulation and real robot experiments demonstrate that our approach significantly outperforms baseline methods and exhibits strong dynamic grasping capabilities.

1. Introduction

Within the broader robotic manipulation pipeline, grasping serves as a pivotal initial step that enables physical interaction. Besides, grasp detection methods have achieved remarkable results by mapping visual inputs, such as images or point-clouds, to viable grasp poses. Recently, visuomotor policies derived from imitation learning [4, 7, 19, 47] have demonstrated significant potential in general-purpose

† Corresponding Author.

robotic manipulation. Trained on large-scale demonstration data, these policies can exhibit zero-shot generalization and achieve robust error correction against external disturbances via closed-loop control. However, specifically for the grasping stage, these general-purpose policies often fall short of specialized grasp detection methods, as modeling the entire action sequence of grasping is an inherently more complex task. Therefore, enhancing the capability of general visuomotor policies to perform fine-grained generalized grasping is a critical research objective.

To address this issue, prior research has primarily focused on two directions: data-centric strategies and the integration of additional prior knowledge. From the data-centric perspective, to alleviate the scarcity of grasping demonstrations, GraspVLA [9] generates a massive dataset Syngrasp-1b of 1 billion simulated frames to train a Vision-Language-Action (VLA) model, which in turn demonstrates remarkable zero-shot sim-to-real transfer capabilities. However, large-scale data generation incurs substantial computational cost: Syngrasp-1b consumes 160 RTX 4090 GPUs for 10 days to simulate, an really uneconomical expense. On the other hand, large VLA models also suffer from high inference latency and low action frequency, which hinders real-time grasping in dynamic scenes. Regarding the integration of additional knowledge, some works [14, 15, 31] enhance imitation learning frameworks to enhance grasp performance by incorporating a grasp detection network, whose accurate predictions of the target grasp pose provide efficient guidance for the policy model. Compared to GraspVLA, these methods can achieve more efficient usage of the demonstration data and reduce inference latency. However, these methods treat grasp poses merely as conditional input for the policy model, which leads to two issues. On the one hand, the input grasp pose is weakly correlated with the output action sequence, making it difficult to provide efficient guidance. On the other hand, the mismatch between the low-semantic grasp pose and the visual inputs causes the policy model to fail to adequately extract information about the spatial distribution of grasps.

To overcome these challenges, drawing inspiration from recent advances in image generation [1, 41, 46], we introduce a novel grasp guidance framework built upon latent diffusion models. In contrast to previous methods, our key innovation is to steer the action generation process by constructing an action latent space and injecting the precise target grasp pose. As illustrated in Fig. 1 (c). Instead of holistically modeling the entire grasp sequence with a single policy, our method disentangles the action latent generation into two distinct components: a target grasp pose and the corresponding motion policy. The former is predicted by a dedicated grasp detection network, while the latter is learned by the diffusion model. This decomposition bridges the gap between the static target grasp pose and the dynamic action sequence by projecting both into a shared latent space. Furthermore, to minimize the mismatch between grasp pose and visual inputs, we provide the latent diffusion policy with a visual graspness cue. This cue is a graspness map [37], which is also generated by the grasp detector, explicitly directs the policy's attention toward grasp-relevant regions.

In this paper, we propose a framework for generalizable grasping policy via latent diffusion following [30]. Under this two-stage framework, we effectively integrate priors including graspness map and grasp pose from grasp detection network. For grasp pose prior, we refine action chunks under the guidance of a grasp pose in latent space encoded by a Variational Auto-Encoder (VAE) in Action Latent Learning stage, enabling more effective steering of low-semantic information. For graspness map prior, we attach the graspness map to the wrist camera image as a visual cue to condition the diffusion model's denoising process in Diffusion on Latent Action Space stage. In each denoising step, we reconstruct the wrist-view image as an auxiliary self-supervised objective to strengthen the policy's conditioning on the graspness cue. During inference, we further propose Heuristic Pose Selector (HPS) which jointly considers grasp pose quality and the current end-effector state to choose the most appropriate grasp pose from candidates as guidance. Experimental results show that our method further improves in-domain grasping success rate by **17.5%** compared to diffusion policy, and yields significant gains in spatial, object and visual generalization of **22.2%**, **46.8%** and **48.3%**, respectively. We also find that our approach remains highly effective in dynamic grasping, demonstrating the practical potential of the proposed method.

2. Related work

Grasp Detection. Grasp detection has been extensively studied over the past decade. This task typically aims to predict feasible grasp poses for objects based on visual observations. GPD [35] pioneers a two-stage pipeline that combined a sampling algorithm in large-scale candidates with a CNN-based scoring module, achieving high precision in 6-DoF grsap detection. Subsequently, end-to-end grasp detection methods [10, 11, 24–27, 37] become a major research focus. [27] proposes a conditional variational grasp generator that models multimodal 6-DoF grasp distributions. GraspNet-1Billion [11] is a large-scale benchmark for general grasping that contains over a billion labeled grasp candidates, greatly alleviating data scarcity and evaluation inconsistency. GSNet [37] introduces the notion of graspness measures the point-wise graspability in the point-cloud. AnyGrasp [10] addresses the spatiotemporal dimension with dense supervision and demonstrates human-like performance on bin-picking and dynamic grasping tasks. With the rise of LLMs and VLMs, works like GraspGPT [34] and

ThinkGrasp [29] have begun to integrate vision–language reasoning to enable task-oriented grasp detection in cluttered scenes. While these open-loop grasp approaches achieve respectable accuracy, the absence of closed-loop perception during grasp execution limits their adaptability and degrades performance in dynamic environments.

Visual Imitation Learning. In recent years, visual imitation learning has advanced substantially. Early behavior cloning works [12, 17] fitted mappings from robot states to actions using expert demonstrations to accomplish robotic manipulation tasks. Subsequent methods such as ACT [47] and Diffusion Policy [7] leverage generative models to produce action chunk from 2D visual observations, achieving significant success. Following [7], diffusion-based models became widely adopted for visual imitation learning. 3D Diffusion Policy [45] and 3D Diffuser Actor [18] extend 2D visual observations to 3D space to enhance spatial perception. [6, 39] and [36, 38] introduce 3D semantic fields and equivariance into diffusion policies respectively to improve sample efficiency and generalization. RDP [42] proposes a visual-tactile diffusion framework that integrates tactile or force information for contact-rich manipulation task. Meanwhile, large VLA models have also begun to adopt diffusion head for action generation after preliminary explorations[19, 48]. Octo [13] uses a diffusion action head to predict action chunking. Then RDT-1B [23] builds a much larger DiT [28] to realize general bimanual manipulation. π_0 [4] and $\pi_{0.5}$ [16] employs flow-matching [21] to achieve higher-frequency and smoother action sequence. Due to the absence of task-specific design for grasping, policy-based approaches are often suboptimal. In contrast, our method effectively addresses the limitations of closed-loop grasping strategies in terms of accuracy and generalization by incorporating prior knowledge of grasping.

Visuomotor Policy for Grasping. In this section, we focus on works that design imitation learning policies for grasping. GraspVLA [9] is a data-centric approach for universal grasping. Fueled by one billion frames of data, it explicitly adopts a Chain-of-Thought (CoT) [40] pipeline named Progressive Action Generation, causing the model to output the bounding box and the grasp pose of the target object before generating the action chunk by flow matching action expert. This pipeline achieves state-of-the-art results in universal closed-loop grasping. Simultaneously, some prior-centric approaches have been proposed that integrate grasp detection module or grasp pose into imitation learning workflows. PPI [43] first leverages discrete key poses (which can be viewed as generalized grasp poses) to guide continuous action generation: the paper treats object pointflow and key gripper poses as intermediate interfaces that steer denoising process, showing effectiveness on long-horizon bimanual tasks. Robograsp [15] injects pose-aware grasping features for planar grasps as one conditioning input to a diffusion policy and synchronously calls a pre-trained grasp detector at every inference timestep; Spatial Robograsp [14] extends this idea to full 6-DoF grasps. GPA-RAM [31] uses pre-trained M2T2 [44] and ResNet as grasp detectors and fuses their implicit features with Spatial Attention Mamba output tokens to predict the next key pose for task execution. In summary, current research has not fully mined the mature grasp detection literature for richer priors—most efforts remain limited to grasp poses—and the methods for introducing those priors still require further exploration.

3. Method

3.1. Overview

Our method can be formulated as learning a visuomotor policy $\pi : \mathcal{O} \rightarrow \mathcal{A}$ that maps observations to action chunks. Our core insight is that grasp priors can not only increase the accuracy of actions in grasping tasks by providing precise grasp configurations and improve generalization by reducing the policy's reliance on raw visual observations and proprioceptive states. To this end, we propose **GraspLDP**, a two-stage trained latent diffusion model as shown in Figure 2. The detailed design rationale and technical specifics are presented in the following subsections.

3.2. Grasp Guidance in Latent Space

Diffusion policy can be formulated as a conditional denoising model conditions on the current timestep observation O_t, which includes RGB images, depth maps, and the robot's proprioceptive states. Concretely, starting from Gaussian noise A_t^k, the denoising network ϵ_θ performs k steps of a Markov process with parameterized Gaussian transitions to produce the final clean action chunk A_t^0:

$$A_t^{k-1} = \alpha \left(A_t^k - \gamma \epsilon_\theta(O_t, A_t^k, k) + \mathcal{N}(0, \sigma^2 \mathbf{I}) \right). \quad (1)$$

To generate actions guided by a target grasp pose $\mathcal{G}$, a natural idea is to include the pose directly as part of the observation or proprioceptive state and condition the denoising process on it. On the one hand, modeling the joint conditional distribution in this way both dilutes the guiding strength of the grasp pose and makes denoising training harder. On the other hand, performing grasp detection and action-chunk denoising sequentially increases inference latency, and low policy latency is itself important for task success. Inspired by prior work [33, 42], we therefore adopt a latent diffusion policy. Concretely, we first use a lightweight VAE encoder to compress action chunks into compact action latents $\mathbf{Z}$:

$$\mathbf{Z} = \mathcal{E}(A). \quad (2)$$

These latents are paired with the trajectory's corresponding grasp pose $\mathcal{G}$ and then reconstructed into action chunks by a asymmetric decoder:

$$\hat{A} = \mathcal{D}(\mathbf{Z} \oplus \mathcal{G}). \quad (3)$$

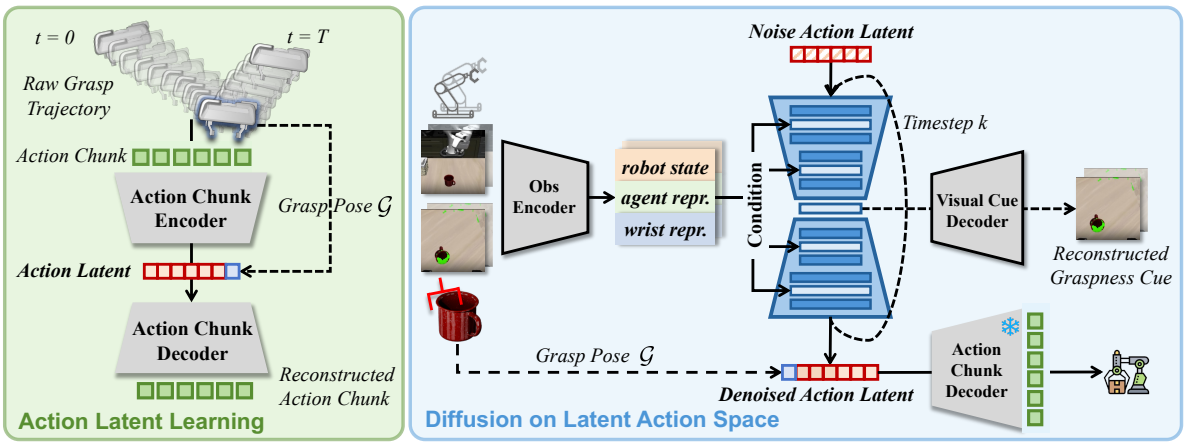

Figure 2. Framework of proposed GraspLDP. In Action Latent Learning stage action chunks are refined under the guidance of a grasp pose in latent space encoded by a VAE. In Diffusion on Latent Action Space stage the graspness cue is used to condition the diffusion model's denoising process and to reconstruct for enhancement.

The VAE is trained using an L2 reconstruction loss with a Kullback-Leibler (KL) penalty loss [20] as follows:

$$\mathcal{L}_{VAE} = \text{MSE}(A, \hat{A}) + \lambda \mathcal{L}_{KL}. \tag{4}$$

Under this scheme, the denoising objective for our latent diffusion policy targets the much more compact action latent representation.

3.3. Visual Graspness Cue

Graspness measures the likelihood that a point in point-clouds affords a feasible grasp. While this notion superficially resembles the concept of affordance [2, 8, 32], graspness is in fact a more precise, geometry-driven form of grasp affordance. As illustrated in Figure 2, we compute point-wise graspness score $s_i \in [0, 1]$ over the depth-projected point-cloud using a pretrained graspness network.

Following the success of previous visual prompting schemes in robotic manipulation, we back-project graspness score to pixel space Ω obtaining graspness map M:

$$\Omega = \{(j, k) \mid 0 \le j < H,\ 0 \le k < W\}, \tag{5}$$

$$M(j, k) = s_{\pi(j,k)},\ (j, k) \in \Omega, \tag{6}$$

where $\pi : \pi(j, k) = i$ is the operator that projects pixels to 3D points. Then the graspness map M is superimposed on the wrist-view RGB image and only points whose graspness exceeds a threshold τ will be counted to mitigate excessive noise contamination to preserve the information content of the original image:

$$O_{cue}(j, k) = \begin{cases} O_{wrist}(j, k), & M(j, k) \le \tau, \\ masked\_color, & M(j, k) > \tau \end{cases}. \tag{7}$$

Then we directly use the resulting masked image as a visual cue to condition the denoising process of policy, aiming to directs the motion towards graspable regions. At the same time, to encourage the model to attend to these visual cues rather than simply rely on condition, we use O_{cue} as an auxiliary self-supervised learning (SSL) objective:

$$\mathcal{L}_{Recon.} = \text{MSE}(O_{cue}, \hat{O}_{cue}), \tag{8}$$

we reconstruct O_{cue} from intermediate representations of the reverse diffusion process and optimize this reconstruction jointly with the diffusion loss:

$$\mathcal{L}_{Diff.} = \text{MSE}\left(\epsilon^k,\ \epsilon_\theta(\bar{\alpha}_k \mathbf{Z}^0 + \bar{\beta}_k \epsilon^k,\ O,\ k)\right), \tag{9}$$

$$\mathcal{L}_{LDP} = \mathcal{L}_{Diff.} + \lambda_{Recon.} \mathcal{L}_{Recon.} \tag{10}$$

3.4. Heuristic Pose Selector

In the inference pipeline, a key challenge is how to select an appropriate from grasp pose candidates prdiected by pretrained grasp detector, since evidently unsuitable pose guidance could degrade the success rate during grasp execution. This process can be formally defined as selecting $\mathcal{G}^*$ from $\mathcal{G} = \{\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_m\}$. We consider two factors in the selection: the score of the grasp predicted by grasp detector and the spatial relationship between the grasp pose and the current end-effector pose. The former represents the intrinsic quality of the grasp, while the latter governs kinematic proximity, yielding smoother and more kinematically feasible end-effector trajectories. Based on these factors we implement a Heuristic Pose Selector (HPS).

First we discard any grasp poses that collide with the environment using collision detection, and then apply non-maximum suppression (NMS) to avoid highly redundant candidate poses. We then score the remaining grasps by their grasp score which is also predicted by grasp detector and keep the top-k candidates, forming the shortlist

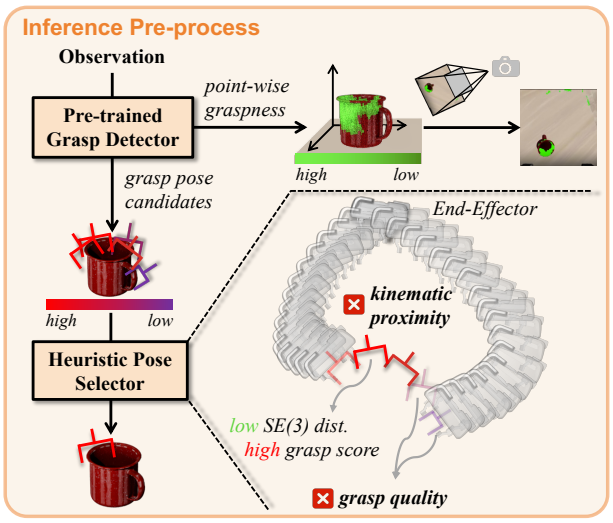

Figure 3. Inference Pre-process presents our inference pipeline with Heuristic Pose Selector.

$\mathcal{G} = \{\mathcal{G}_{m0}, \mathcal{G}_{m1}, \ldots, \mathcal{G}_{mk}\}$. Given the current end-effector pose P, we measure the pose error between P and each candidate $\mathcal{G}_j$ using a pseudo-metric called SE(3) geodesic distance d since there is no bi-invariant metric on SE(3) space [3, 5]:

$$\xi = \log(P^{-1}\mathcal{G}_j),\ j \in \{m_0, m_1, \ldots, m_k\}, \tag{11}$$

$$d_{\mathcal{G}_j,W} = \sqrt{\xi^\top W\, \xi}. \tag{12}$$

Here $W = \mathrm{diag}(w_t, w_t, w_t, w_r, w_r, w_r)$ is a 6×6 diagonal weighting matrix used to make rotation and translation commensurate. We then select the grasp $\mathcal{G}^*$ that minimizes this distance d_j as our final grasp choice:

$$\mathcal{G}^* = \underset{\mathcal{G}_j \in \mathcal{G}}{\arg\min}\ d(\mathcal{G}_j). \tag{13}$$

With the selected grasp and denoised action latent, the VAE decoder reconstructs the final action chunk as Equation 3, which will be excluded by manipulator.

4. Experiments

4.1. Experimental Setup

Benchmark. We conduct simulation experiments on the LIBERO [22] benchmark covering data collection, training, and evaluation. We curate and filter a training set of roughly **12K high-quality demonstrations** over **20 objects**, covering varied object poses and diverse grasp poses. For evaluation of generalization we select unseen objects from both the similar and novel splits in Graspnet-1billion [11], which are held-out not only for our GraspLDP but also for the pre-trained grasp detection network used in the pipeline to ensure a fair comparison. More detailed procedures for constructing the benchmark are provided in the Appendix.

Baselines. To validate the architectural advances of our GraspLDP, we compare to vanilla Diffusion Policy [7], and to the generalist manipulation policy OpenVLA [19]—both in zero-shot setting and in a version fine-tuned on our dataset. Since our work focuses specifically on grasping, we also compare with the state-of-the-art method GraspVLA [9]. In addition, we define an Ours Baseline like prior studies [14, 15] which treats the grasp pose merely as another conditioning input concatenated with other observations during the denoising process and doesn't use our two-stage latent design.

Metric. In all the following experiments, **Success Rate (SR)** is defined as the percentage of trails that successfully complete the grasping task out of the total number of trails with the maximum time-step threshold $T_{\max} = 150$. A grasp is counted as successful when the target object is stably gripped and lifted above a specified height. For cluttered scenarios evaluation, **Scene Completion Rate (SCR)** is adopted for evaluating the percentage of how many objects of the scene successfully grasped within the allowed attempts. To quantify how well our method follows the target grasp pose guidance T_{GP}, we introduce a new metric, **Grasp Frame Error (GFE)**. For each trajectory, we denote the pose of frame at which the gripper begins to close as Grasp Frame Pose (GFP). For $T_{GFP}, T_{GP} \in SE(3)$ we follow Equations (11) and (12) and compute the GFE as defined there.

$$\textbf{GFE} = \sqrt{\log(T_{GFP}^{-1} T_{GP})^\top\, W\, \log(T_{GFP}^{-1} T_{GP})}. \tag{14}$$

4.2. Simulation Evaluation

In Domain Evaluation. To evaluate the capacity of our policy, we first conduct in-domain experiments where both the objects and their poses appear in the training set. For each test object we evaluate five distinct object poses and repeat each pose 13 trials, which ensures to capture the diversity of possible grasping strategies and the large number of repetitions increases statistical reliability.

As shown in Table 1, the Diffusion Policy achieves a relatively good grasp success rate in the in-domain test set, while the fine-tuned OpenVLA still achieves only 57.5% SR. GraspVLA demonstrates strong zero-shot grasping ability, but with high variance: it achieves nearly 100% on some objects while nearly 0% on others. In contrast, GraspLDP attains the highest SR of **80.3%**. By effectively incorporating grasp priors, our policy achieves greater grasping precision as shown in Figure 5, whereas pervious methods often produce imprecise grasp frame poses that lead to collisions or outright grasp failures.

Generalization Evaluation. We evaluate generalization along three axes including **Spatial Generalization**, **Object Generalization**, and **Visual Generalization**. As summarized in Table 1, the Diffusion Policy suffers a large performance drop when confronted with novel objects or light

Table 1. Results of evaluation in simulator. In Domain denotes cases where both the objects and their poses were present in the training data; Spatial Generalization measures how well the model handles those training objects placed in unseen poses; Object Generalization assesses performance on entirely novel objects; and Visual Generalization tests robustness under visual disturbances like lighting changes. The † refers to the model that has been fine-tuned on our dataset for fair comparison.

Method	In Domain	Spatial Generalization	Object Generalization	Visual Generalization	Average
Diffusion Policy [7]	62.8 (204/325)	48.9 (159/325)	11.4 (37/325)	16.3 (53/325)	34.9
OpenVLA [19]	1.2 (4/325)	0.9 (3/325)	1.5 (5/325)	0.0 (0/325)	0.9
OpenVLA† [19]	57.5 (187/325)	41.2 (134/325)	14.5 (47/325)	12.3 (40/325)	31.4
GraspVLA [9]	50.8 (165/325)	49.5 (161/325)	46.8 (152/325)	51.7 (168/325)	49.7
Ours Baseline	72.3 (235/325)	59.1 (192/325)	48.3 (157/325)	47.7 (155/325)	56.9
GraspLDP	**80.3 (261/325)**	**71.1 (231/325)**	**58.2 (189/325)**	**64.6 (210/325)**	**68.6**

Table 2. Results of ablation study. ID, SG, OG, and VG denote In Domain, Spatial, Object and Visual Generalization, respectively. GC and LG denotes Graspness Cue and Latent Guidance. CG denotes Condition Guidance used in Ours Baseline.

Method	ID		SG		OG		VG	
	SR↑	GFE↓	SR↑	GFE↓	SR↑	GFE↓	SR↑	GFE↓
GraspLDP	**80.3**	**1.33**	**71.1**	**1.61**	**58.2**	**2.10**	**64.6**	**1.92**
w/o GC	77.4 (-2.9)	1.49 (+0.16)	67.3 (-3.8)	1.81 (+0.20)	54.2 (-4.0)	2.33 (+0.23)	57.5 (-7.1)	2.30 (+0.38)
w/o LG w/ CG	73.5 (-6.8)	1.58 (+0.25)	62.2 (-8.9)	2.07 (+0.46)	52.3 (-5.9)	2.42 (+0.32)	54.5 (-10.1)	2.35 (+0.43)
w/o LG	60.6 (-19.7)	-	49.8 (-21.3)	-	21.2 (-37.0)	-	19.4 (-45.2)	-
w/o GC & LG	55.1 (-25.2)	-	46.2 (-24.9)	-	16.0 (-42.2)	-	15.7 (-48.9)	-

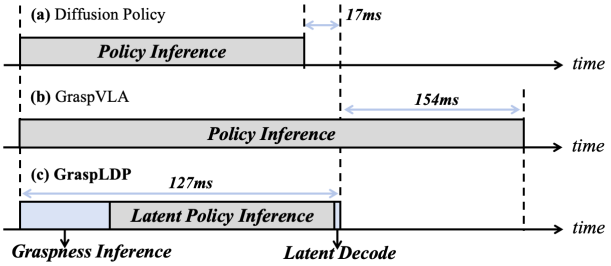

Figure 4. Inference latency of three methods on an RTX 4090 GPU, with the policy action horizon aligned to 8 for each inference. Results of GraspVLA are after acceleration with $torch.compile()$.

Table 3. Ablation study on selection strategy of grasp pose.

Method	ID	SG	OG	VG
w/ random	66.8 (-13.5)	60.6 (-10.5)	51.7 (-6.5)	54.5 (-10.1)
w/ highest	72.0 (-8.3)	63.1 (-8.0)	55.4 (-2.8)	58.2 (-6.4)
w/ nearest	69.5 (-10.8)	61.8 (-9.3)	53.8 (-4.4)	56.0 (-8.6)
w/ HPS	**80.3**	**71.1**	**58.2**	**64.6**

conditions: out-of-distribution (OOD) visual observations substantially degrade its capability. By contrast, GraspVLA shows strong zero-shot generalization and stable performance across evaluation splits, which we attribute to its model capacity and large-scale simulation data. Our method achieves the best results across all three generalization settings, indicating that the graspness cue and grasp pose guidance enable the model to remain robust under OOD visual observations and still generate correct grasp trajectories.

Inference Time Analysis. Because our method adds extra processing at each inference timestep, a quantitative analysis of inference latency is necessary. As shown in Figure 4, our method introduces two additional components: graspness inference ($36\ ms$) and latent decode ($< 1ms$) compared to vanilla Diffusion Policy. At the same time, a smaller dimensionality of latent space leads to faster policy inference. Overall, GraspLDP is only about 15% slower than diffusion policy with the same configuration, delivering nearly a twofold improvement in overall success rate. For GraspVLA, its inference latency remains $154ms$ higher than GraspLDP even with acceleration, which means our method can respond faster in dynamic scene.

4.3. Ablation Study

We first ablate the major components of the framework to evaluate their individual contributions. **Graspness Cue** and **Latent Guidance** refer to two incremental augmentations: the former adds the geometry-driven graspness visual cue together with an auxiliary self-supervised reconstruction objective, while the latter conducts grasp pose guid-

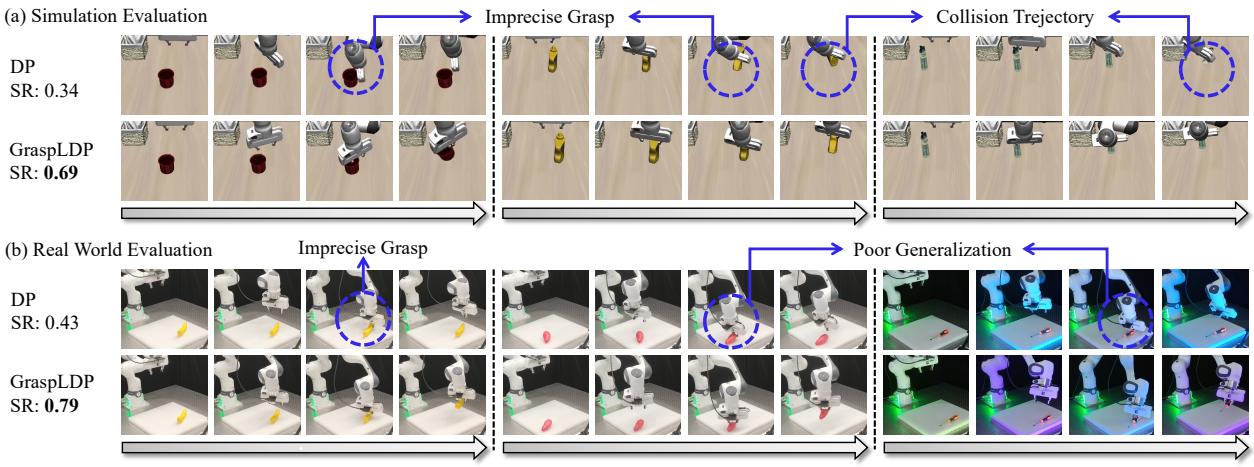

Figure 5. Qualitative experimental analysis. (a) Grasping trials using objects "mug", "mustard bottle", and "thera med" in simulator. (b) Real world grasping trials corresponding to in domain, object generation, and visual generation performance. In particular, we use colored LED strips in low-light conditions to simulate visual interference.

ance inside the latent space of the action chunk. We also report **GFE** to quantify how well each method follows the guidance of target grasp pose. As shown in Table 2, the Graspness Cue increases overall grasp success rate, with the effect most pronounced in VG evaluation split. We attribute this to the geometric, illumination-invariant nature of graspness: under challenging lighting or visual noise, the graspness cue provides a lighting-robust visual hint that draws the end-effector toward higher graspness regions. Large reduction on SR when we remove LG and increase on GFE when we relpace LG with CG indicate that guiding the policy in the compact latent action representation is a more effective and precise way to incorporate the grasp pose prior.

Additionally, we perform an ablation study on our proposed grasp pose selection strategy. Using the fully trained GraspLDP model, we replace the HPS at inference time with three simple alternatives: **random** (uniformly sample a candidate from $\mathcal{G}$), **highest** (pick the grasp with the largest grasp score), and **nearest** (pick the grasp with the smallest SE(3) distance to the current end-effector pose). Table 3 reports the results. The experiments clearly demonstrate the benefit of HPS's balanced design. HPS jointly trades off grasp quality and kinematic proximity, yielding smoother, more feasible end-effector trajectories and significantly better task success than any single-criterion selection rule.

4.4. Real World Evaluation

Settings. Our real world evaluation setup is shown in Figure 6. We perform experiments on a Franka Research 3 manipulator. The wrist-view camera is an Intel RealSense D405, the agent-view camera is an Intel RealSense L515. We include an additional side-view RealSense D405 solely for evaluating GraspVLA. The workspace covers a $50 \times 50 \times 20\ cm^3$ region. Our object set consists of everyday items: 10 training objects and 13 test objects, including most rigid bodies and a small number of articulated and deformable objects. For each training object we collect 50 demonstrations (500 demonstrations total) to train both our method and Diffusion Policy; we also compare against GraspVLA and AnyGrasp[10]. To ensure fair evaluation, the robot's initial pose, workspace bounds and all camera extrinsics remain fixed across every experiment. Each evaluation set contains 10 objects, and for each object we run 10 trials with different poses sampled across the workspace.

Figure 6. Settings of real world evaluation and daily objects used for training and evaluation.

Table 4. Results of real world evaluation. Because it's difficult to ensure identical object poses in the real world, we merge the ID and SG splits into a single evaluation set.

Method	ID&SG	OG	VG	Avg
Diffusion Policy [7]	65.0	43.0	21.0	43.0
GraspVLA [9]	29.0	27.0	22.0	26.0
Anygrasp [10]	82.0	**78.0**	**79.0**	**79.7**
GraspLDP	**84.0**	75.0	77.0	78.7

Table 5. Results of cluttered scenarios evaluation in real world.

Method	Scene1		Scene2		Scene3		Scene4		Avg	
	SR↑	SCR↑	SR↑	SCR↑	SR↑	SCR↑	SR↑	SCR↑	SR↑	SCR↑
Diffusion Policy [7]	10.0	20.0	0.0	0.0	20.0	28.6	10.0	12.5	10.0	15.4
GraspVLA [9]	20.0	40.0	0.0	0.0	30.0	42.9	10.0	12.5	15.0	23.1
Anygrasp [10]	**83.3**	**100.0**	66.7	**100.0**	77.8	**100.0**	**60.0**	**75.0**	70.1	**92.3**
GraspLDP	**83.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**60.0**	**75.0**	**82.8**	**92.3**

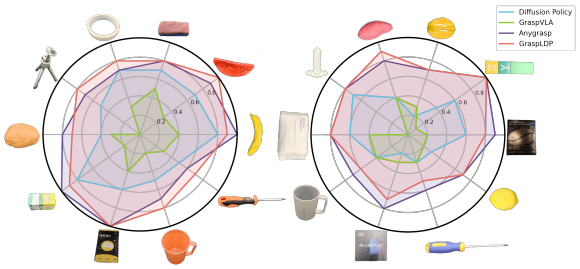

Figure 7. Detailed results of in domain&spatial generalization and object generalization in the real world.

Main Results. The overall experimental results are summarized in Table 4. Our method achieves the best success rate of **84.0%** on the ID&SG evaluation, and it maintains strong robustness under novel objects and extreme visual variations—achieving an overall SR comparable to AnyGrasp. This indicates that our design enables the closed-loop policy to inherit the grasp detector's strong generalization to novel objects and visual changes, significantly outperforming the Diffusion Policy and GraspVLA. Figure 7 provides a detailed per-object breakdown of each method's performance for the ID&SG and OG evaluations.

Cluttered Scenarios Evaluation. We conduct grasping experiments in cluttered scenes using the same real robot setup. As shown in Appendix, we design four cluttered scenarios containing from 5 to 8 objects; Scene 4 even includes a more challenging stacked-object configuration. The progressively increasing difficulty makes it clearer to reveal each method's limits in handling cluttered Scenarios.

The results are summarized in Table 5, both our method and AnyGrasp achieve the highest SCR of **92.3%**, while our method attains a **12.7%** higher SR. This outcome is notable because, unlike AnyGrasp which is trained on multi-object point-cloud data, GraspLDP was trained only on single object grasping demos. GraspLDP successfully grasps every object in Scene 1–3 and still demonstrates strong performance in the stacking scenario, indicating good generalization across the height of workspace which is often overlooked in previous work. Due to zero-shot deployment, GraspVLA could only complete the simplest grasps, though it still outperforms the diffusion policy which is frequently confused by multi-object interference.

Table 6. Results of dynamic grasp task.

Method	banana moving	watermelon moving	mug handover
Diffusion Policy [7]	✗/✗/✗	✗/✗/✗	✓/✗/✗
GraspVLA [9]	✓/✓/✗	✓/✗/✗	✗/✗/✗
Anygrasp [10]	✓/✓/✗	✓/✗/✗	✓/✗/✗
GraspLDP	✓/✓/✗	✓/✓/✓	✓/✓/✗

Dynamic Grasp. We also validate our method's dynamic grasping capability on the same real robot setup. For GraspLDP and Diffusion Policy we shorten the action horizon from 8 to 4 so the controller can react more promptly to object motion. Results in Table 6 show that the diffusion policy trained on static grasp data almost fails to adapt to dynamic scenes. By contrast, our method, which updates the pose of guiding grasp synchronously, can track, approach, and grasp moving objects. With a higher inference frequency, it achieves markedly better performance than GraspVLA. Compared to AnyGrasp, which uses a tracker to maintain temporal consistency across adjacent frames, rapid changes in the target pose still induce abrupt changes in end-effector trajectories and cause failures, our HPS explicitly accounts for the SE(3) distance between the current end-effector pose and grasp pose candidates, and the policy itself conditions on recent end-effector states. As a result, GraspLDP generates more continuous and smoother grasp trajectories and attains higher success rates.

5. Conclusion

In this work, we propose a framework of generalizable latent diffusion policy for grasping. **GraspLDP** adopt grasp pose guidance in latent action space and leverages auxiliary self-supervised reconstruction of graspness cues to achieve higher grasping precision and improve spatial, visual and object generalization. It can handle cluttered multi-object scenarios and shows promising capability in dynamic scenarios. The method demonstrates strong performance and scalability, offering a promising foundation for future work on robotic foundation model for grasping and manipulation.

Limitations. Our method may still face challenges when

handling highly deformable or fragile objects such as egg or beaker. In future works we plan to incorporate high frequency tactile and force/torque signals into the framework and explore more general grasping policy.

Acknowledgment

This work is partly supported by the National Key Research and Development Plan (2024YFB3309302).

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):1–11, 2023. 2

[2] Daniel Baldauf and Heiner Deubel. Attentional landscapes in reaching and grasping. *Vision Research*, 50(11):999–1013, 2010. 4

[3] Timothy D Barfoot and Paul T Furgale. Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Transactions on Robotics*, 30(3):679–693, 2014. 5

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 3

[5] José Luis Blanco-Claraco. A tutorial on se(3) transformation parameterizations and on-manifold optimization. *arXiv preprint arXiv:2103.15980*, 2021. 5

[6] Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, et al. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1735–1744, 2025. 3

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023. 1, 3, 5, 6, 7, 8

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4

[9] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, Zhizheng Zhang, and He Wang. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. In *Conference on Robot Learning*, 2025. 2, 3, 5, 6, 7, 8

[10] Haoshu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023. 1, 2, 7, 8

[11] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 5

[12] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, 2022. 3

[13] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, 2024. 3

[14] Yiqi Huang, Travis Davies, Jiahuan Yan, Xiang Chen, Yu Tian, and Luhui Hu. Robograsp: A universal grasping policy for robust robotic control. *arXiv preprint arXiv:2502.03072*, 2025. 2, 3, 5

[15] Yiqi Huang, Travis Davies, Jiahuan Yan, Jiankai Sun, Xiang Chen, and Luhui Hu. Spatial robograsp: Generalized robotic grasping control policy. *arXiv preprint arXiv:2505.20814*, 2025. 2, 3, 5

[16] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Conference on Robot Learning*, 2025. 3

[17] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, 2022. 3

[18] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *Robotics: Science and Systems*, 2024. 3

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024. 1, 3, 5, 6

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. 4

[21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 3

[22] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, 2023. 5

[23] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, 2025. 3

[24] Haoxiang Ma and Di Huang. Towards scale balanced 6-dof grasp detection in cluttered scenes. In *Conference on robot learning*, 2022. 2

[25] Haoxiang Ma, Ran Qin, Modi Shi, Boyang Gao, and Di Huang. Sim-to-real grasp detection with global-to-local rgb-d adaptation. In *IEEE International Conference on Robotics and Automation*, 2024.

[26] Haoxiang Ma, Modi Shi, Boyang Gao, and Di Huang. Generalizing 6-dof grasp detection via domain prior knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[27] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023. 3

[29] Yaoyao Qian, Xupeng Zhu, Ondrej Biza, Shuo Jiang, Linfeng Zhao, Haojie Huang, Yu Qi, and Robert Platt. Thinkgrasp: A vision-language system for strategic part grasping in clutter. In *Conference on Robot Learning*, 2025. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[31] Juyi Sheng, Yangjun Liu, Sheng Xu, Zhixin Yang, and Mengyuan Liu. GPA-RAM: grasp-pretraining augmented robotic attention mamba for spatial task learning. *arXiv preprint arXiv:2504.19683*, 2025. 2, 3

[32] Hyun Oh Song, Mario Fritz, Daniel Goehring, and Trevor Darrell. Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering*, 13 (2):798–809, 2015. 4

[33] Wenhui Tan, Bei Liu, Junbo Zhang, Ruihua Song, and Jianlong Fu. Rold: Robot latent diffusion for multi-task policy modeling. *International Conference on Multimedia Modeling*, 2025. 3

[34] Chao Tang, Dehao Huang, Wenqi Ge, Weiyu Liu, and Hong Zhang. Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping. *IEEE Robotics and Automation Letters*, 8(11):7551–7558, 2023. 2

[35] Andreas Ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *International Journal of Robotics Research*, 36(13-14):1455–1473, 2017. 2

[36] Chenrui Tie, Yue Chen, Ruihai Wu, Boxuan Dong, Zeyi Li, Chongkai Gao, and Hao Dong. Et-seed: Efficient trajectory-level se (3) equivariant diffusion policy. In *International Conference on Learning Representations*, 2025. 3

[37] Chenxi Wang, Haoshu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2

[38] Dian Wang, Stephen Hart, David Surovik, Tarik Kelestemur, Haojie Huang, Haibo Zhao, Mark Yeatman, Jiuguang Wang, Robin Walters, and Robert Platt. Equivariant diffusion policy. In *Conference on Robot Learning*, 2024. 3

[39] Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kelestemur, Jiuguang Wang, and Yunzhu Li. Gendp: 3d semantic fields for category-level generalizable diffusion policy. In *Conference on Robot Learning*, 2024. 3

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 3

[41] Zilong Wu, Hideki Murata, Nayu Takahashi, Qiyu Wu, and Yoshimasa Tsuruoka. Latentps: Image editing using latent representations in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 2

[42] Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. In *Robotics: Science and Systems*, 2025. 3

[43] Yuyin Yang, Zetao Cai, Yang Tian, Jia Zeng, and Jiangmiao Pang. Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation. In *Robotics: Science and Systems*, 2025. 3

[44] Wentao Yuan, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. M2t2: Multi-task masked transformer for object-centric pick and place. In *Conference on Robot Learning*, 2023. 3

[45] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *Robotics: Science and Systems*, 2024. 3

[46] Yu Zeng, Yang Zhang, Liu Jiachen, Linlin Shen, Kaijun Deng, Weizhao He, and Jinbao Wang. Hairdiffusion: Vivid multi-colored hair editing via latent diffusion. In *Advances in Neural Information Processing Systems*, 2024. 2

[47] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023. 1, 3

[48] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2023. 3