Raise One and Infer Three: Toward Reasoning- and Memory-Augmented Diffusion Policy Generalization

Yihang Zhu, **Yuxuan Wang**, **Tong Li**, **Jiexi Yan**, **Xu Yang** and **Cheng Deng***
Xidian University, China
{zhuyihang, wangyuxuan, litong1}@stu.xidian.edu.cn, {jxyan1995, xuyang.xd, chdeng.xd}@gmail.com

Abstract

Diffusion policy has shown impressive performance in robotic manipulation tasks while struggling with out-of-distribution shifts and limited demonstrations. Recent advances primarily focus on improving geometric or perceptual representations for diffusion policy. However, these approaches rely heavily on instantaneous observations, making them vulnerable when visual inputs deviate from the training distribution. Our key insight is that robust generalization in diffusion policy requires both structured reasoning over skill compositions and an explicit memory mechanism for accumulating and reusing learned policy knowledge. To this end, we propose "raise one and infer three" diffusion policy (ROITDP), a novel approach that introduces two complementary mechanisms. First, we propose a reasoning mechanism built upon the Chain-of-Skill Noise Watermark, which encodes skill-level reasoning graph representations into the initial noise space, thereby enabling temporally coherent multi-step reasoning throughout the diffusion process under distribution shifts. We further introduce a memory mechanism in the form of a Self-evolving Policy Knowledge Bank that discretizes spatio-temporally refined trajectories into reusable skill primitives via a VQ-VAE architecture, providing adaptive memory to guide and refine action generation. Extensive experimental results on both simulated and real-world environments demonstrate the superiority and robustness of our method.

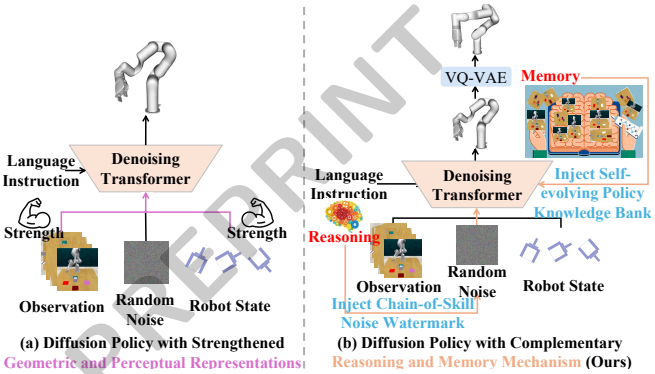

Figure 1: Comparison of diffusion policy learning paradigms for manipulation. **Left:** Conventional diffusion policies primarily improve generalization by strengthening geometric and perceptual representations, but rely heavily on instantaneous visual observations, which can be unreliable under out-of-distribution scenarios where visual inputs are corrupted or shifted. **Right:** We propose ROITDP, which introduces complementary reasoning and memory mechanisms by injecting skill-level reasoning into the noise space and maintaining an adaptive policy knowledge bank, enabling diffusion policies to "raise one and infer three" under distribution shifts.

1 Introduction

Imitation learning offers a powerful paradigm for enabling robots to acquire diverse actions by mimicking expert demonstrations without the need to manually design reward functions or perform extensive trial-and-error exploration. It has attracted significant attention and found broad applications in robotics [Haldar *et al.*, 2023; Jiang *et al.*, 2025], autonomous driving [Hu *et al.*, 2023; Cheng *et al.*, 2024] and legged locomotion [Li *et al.*, 2023; Tang *et al.*, 2024]. Recently, Diffusion Policy [Chi *et al.*, 2023] shows strong potential for imitation learning by modeling complex and multimodal action distributions in an iterative denoising framework. However, due to the high cost of collecting expert demonstrations and the increasing demand for robust performance in unseen scenarios, a fundamental challenge remains: how to enable diffusion-based imitation learning algorithms to acquire generalizable skills from limited demonstrations.

A growing body of work has attempted to improve the generalization of diffusion-based policies by strengthening their geometric and perceptual representations. For instance, [Zhu *et al.*, 2025a] incorporate $\mathrm{SE}(3)$-equivariant structures into diffusion policies to better handle geometric variations. Simultaneously, [Wang *et al.*, 2024] leverages 3D semantic fields constructed from multi-view observations to further improve its generalization ability. Additionally, [Ze *et al.*, 2025] departs from purely 2D representations and performs action diffusion directly in a 3D scene space, which allows

*Corresponding author.

the policy to jointly model geometry, perception, and action throughout the training process. Despite these advances, such approaches still exhibit some limitations.

First, **infer three but no reasoning.** Most existing works primarily rely on implicit correlations of data, while overlooking explicit skill-level reasoning over task. Among these methods, language instructions are usually regarded as static condition signals, without modeling the step-by-step logic required for completing complex tasks. As a consequence, these methods struggle to perform multi-step reasoning, and their performance degrades significantly in tasks that require precise temporal ordering and long-horizon planning. Furthermore, **raise one but no memory.** Current diffusion-based policies also lack an explicit mechanism for retaining and reusing limited policy knowledge. During the denoising process, each step is driven by the current observations, with little access to historical context and accumulated policy knowledge. Consequently, Diffusion Policy cannot effectively guide action generation during denoising, nor can it leverage historical experience to refine drifting trajectories, leading to unstable behavior under distribution shifts.

To address these limitations, as shown in Fig. 1, we explore an integration of reasoning and memory mechanisms into diffusion policy. To be specific, we propose ROITDP, a novel architecture that explicitly models reasoning- and memory-aware skill characteristics. ROITDP consists of two key modules: Chain-of-Skill Noise Watermark, which injects reasoning-aware skill structure into the diffusion process. Given a language instruction, a large language model first decomposes it into an ordered sequence of skill-level sub-goals. However, a flat sequence alone is insufficient to capture the temporal dependencies and logical constraints required for long-horizon decision making. Therefore, we further organize these sub-goals into a Language Reasoning Graph, which encodes the structured relationships among skills and provides a hierarchical representation of skill-level reasoning. Motivated by the semantic neutrality and temporal globality of the noise space, which render it robust to visual domain shifts, we project the skill-level graph representation into the initial noise space as a noise watermark to guide the denoising process in a temporally coherent manner; and the Self-evolving Policy Knowledge Bank, which leverages the ability of VQ-VAE to preserve essential data semantics in a compact latent space and discretizes denoised action trajectories refined through spatio-temporal visual cues into reusable and generalizable skill primitives. These discrete skill representations constitute an abstract set of policy knowledge units distilled from historical experience, which can be reused to guide and refine action generation. Extensive experiments demonstrate that our method achieves state-of-the-art performance, particularly excelling in out-of-distribution scenarios. Our chief contributions are threefold:

- We explore two novel mechanisms, i.e., skill-level reasoning and memory mechanisms, for diffusion policy generalization with few demonstrations.
- We propose a novel framework ROITDP, comprising two core modules: the Chain-of-Skill Noise Watermark for injecting reasoning-aware skill structure into the diffusion process and the Self-evolving Policy Knowledge Bank for capturing reusable and generalizable skill primitives from historical experience.
- We achieve new state-of-the-art performance in a broad range of simulated and real-world robotic manipulation tasks, demonstrating superior accuracy and robustness under diverse environmental perturbations.

2 Related Work

Diffusion-based Imitation Learning. As one of the earliest works to introduce diffusion models into imitation learning, Diffusion Policy [Chi *et al.*, 2023] formulates behavior cloning as a conditional denoising process over action distributions, demonstrating strong capability in modeling multimodal behaviors and providing stable training. Despite these advantages, Diffusion Policy still suffers from limited generalization to out-of-distribution scenarios. Subsequent works enhance generalization by incorporating richer geometric and perceptual representations, including point cloud-based spatial modeling in [Ze *et al.*, 2024; Ze *et al.*, 2025], $\mathrm{SIM}(3)$-equivariant neural architectures in [Yang *et al.*, 2024; Zhu *et al.*, 2025a], and multi-view 3D semantic fields in [Wang *et al.*, 2024; Yan *et al.*, 2025]. Other approaches, e.g., [Xu *et al.*, 2024], integrate self-supervised 2D optical flow estimation into a diffusion-based policy framework. Despite these advances, most diffusion-based models still rely heavily on implicit correlations from expert demonstrations, lacking explicit mechanisms for structured reasoning and long-term knowledge retention, which leads to performance degradation under distribution shifts. In contrast, our work explicitly integrates skill-level reasoning and adaptive policy memory into diffusion-based policy learning, allowing our model to generalize more effectively to unseen scenarios.

Generalizable Manipulation for Robotics. Generalizable robotic manipulation remains a fundamental challenge in imitation learning. One prominent direction explores learning manipulation directly from demonstrations via end-to-end visuomotor policies. Transformer-based models [Brohan *et al.*, 2022; Shridhar *et al.*, 2023; Goyal *et al.*, 2023; Li *et al.*, 2024] predict robot actions from visual and language inputs, but typically require large numbers of demonstrations to generalize effectively across diverse scenes. To better capture 3D structure, methods such as [Lyu *et al.*, 2025], [Zhu *et al.*, 2025b], and [Wang *et al.*, 2025] leverage voxelized or multi-view 3D representations, improving spatial reasoning at the expense of increased computational complexity. Our method advances these developments by introducing both a skill-chain noise watermark and a self-evolving policy knowledge bank, which are reusable, adaptive, and compact, enabling efficient reuse of learned skills and robust long-horizon manipulation under distribution shifts.

3 Overview of Diffusion Policy

Based on the Denoising Diffusion Probabilistic Model (DDPM), the Diffusion Policy is designed to model multimodal distributions in behavior cloning. This method learns

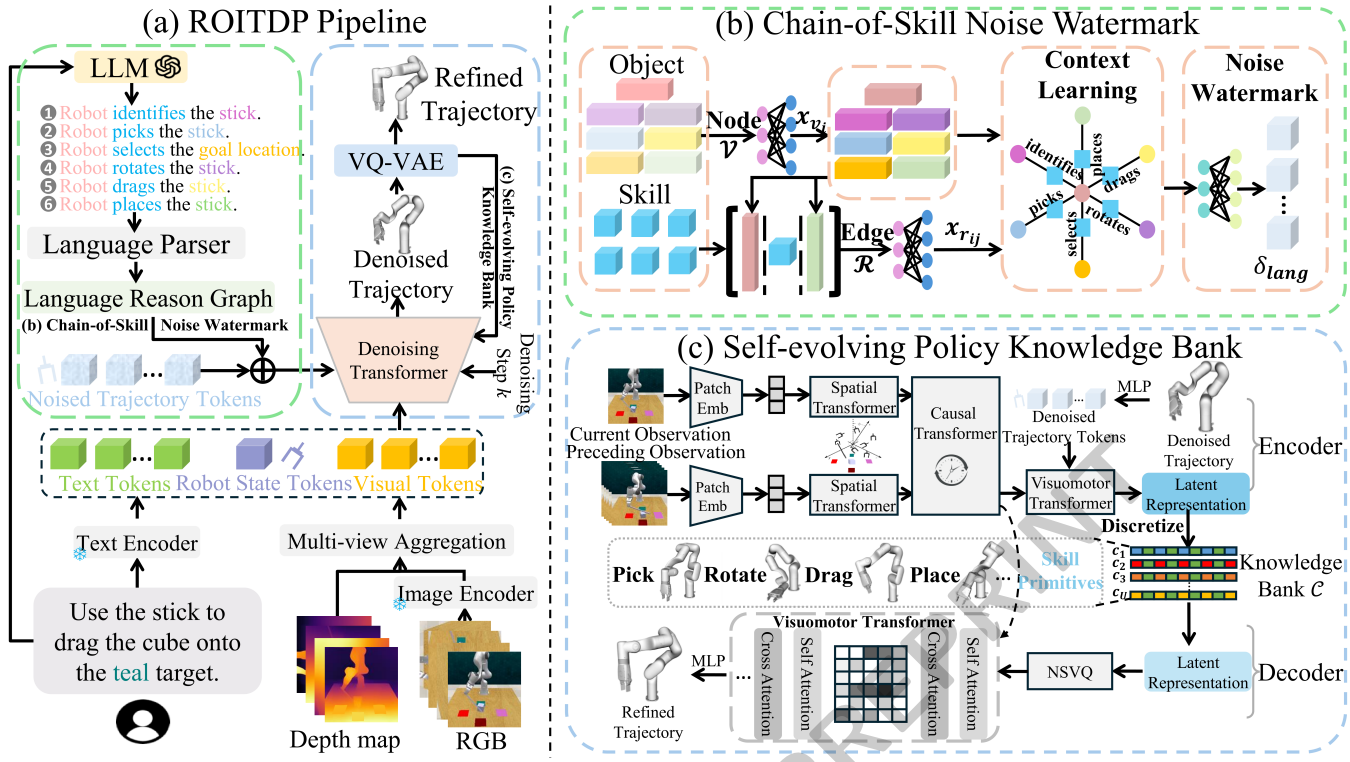

Figure 2: ROITDP for "raise one and infer three" diffusion policy. (a) Overall pipeline. Given language instructions, visual observations, and robot states, ROITDP performs diffusion-based action generation conditioned on both reasoning and memory mechanisms. (b) Chain-of-Skill Noise Watermark. Language instructions are decomposed into skill-level sub-goals and organized as a Language Reasoning Graph, whose structured semantics are injected into the initial noise space to guide temporally coherent denoising. (c) Self-evolving Policy Knowledge Bank. Spatio-temporally refined trajectories are discretized via a VQ-VAE into reusable skill primitives, forming an adaptive policy memory that supports long-horizon reasoning and robust action generation.

a noise prediction function $\epsilon_\theta(O, A + \epsilon^k, k)$ through a neural network to predict the noise component ϵ^k. During training, the model samples observation-action pairs (O, A) from the expert dataset and injects random noise ϵ^k, which has the same dimensionality as A, into the action. The training objective is formulated as:

$$\mathcal{L} = \text{MSE}\left(\epsilon^k, \epsilon_\theta(O, A + \epsilon^k, k)\right). \quad (1)$$

During inference, Diffusion Policy starts from a randomly sampled noisy action $A^k \sim \mathcal{N}(0, 1)$ and generates the final action A_0 through an iterative denoising process of K steps:

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

where α, γ and σ are functions of denoising step k and depend on the noise scheduler. However, as a data-driven posterior learning method, the Diffusion Policy relies excessively on the current observation O for decision-making. When encountering out-of-distribution scenarios O' (e.g., variations in lighting, texture, or object geometry), the model may suffer from policy shift [Zhang *et al.*, 2024], resulting in substantial performance degradation.

4 ROITDP for Diffusion Policy

In this paper, we propose a new ROITDP architecture, aiming to enable Diffusion Policy to "raise one and infer three" by jointly leveraging reasoning and memory mechanisms, thereby enhancing performance and robustness.

Specifically, in Fig. 2 (a), we follow the work [Ke *et al.*, 2024] and represent all conditioning signals as a unified set of tokens. The noised trajectory and robot state are jointly encoded as 3D tokens by mapping noisy poses and proprioceptive states to latent embeddings paired with their corresponding 3D positions. Visual observations, consisting of one or more posed RGB-D images, are converted into 3D visual tokens via CLIP-based patch feature extraction and depth-based back-projection, with multi-view tokens aggregated when available. Language instructions are encoded as text tokens using a pre-trained CLIP text encoder.

In practice, the generalization of Diffusion Policy hinges on two complementary cues. One cue is to explicitly model the reasoning structure of actions, i.e., capturing temporally ordered skill chains that support multi-step decision making. The other cue is to sufficiently leverage historical policy memory, i.e., distilling reusable skill primitives from past experience to guide the denoising process and correct the denoised trajectory. With this in mind, we separately design a chain-of-skill noise watermark and self-evolving policy knowledge bank to effectively capture skill-related reasoning and memory characteristics within diffusion-based policy.

4.1 Chain-of-Skill Noise Watermark

Task Planning with LLM. Diffusion Policy often encounters tasks that require complex multi-step reasoning. In such cases, a simple instruction L alone may not be sufficient to convey the detailed action logic needed to accomplish the task. To address this, as illustrated in Fig. 2 (a), we employ the Large Language Model (LLM) to decompose the instruction into a structured skill-level reasoning chain. Rather than directly planning executable actions, the Chain-of-Skill provides an explicit semantic decomposition of the task into ordered skill-level sub-goals (e.g., "approach the object," "grasp the object," and "move to the target"), thereby capturing the latent structure of the task.

Language Reasoning Graph. In general, LLM-generated Chain-of-Skill sequences are free-form texts that possess a context-dependent nature. To further enhance the model's understanding of the semantic reasoning underlying each substep, we construct a Language Reasoning Graph based on the Chain-of-Skill output.

Particularly, as illustrated in Fig. 2 (b), we employ an off-the-shelf graph parser [Schuster *et al.*, 2015] to perform syntactic dependency parsing and build a language graph $\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$. The graph consists of a set of nodes $\mathcal{V} = \{v_i\}$ and edges $\mathcal{R} = \{r_{ij}\}$, where each node represents an object mentioned in the Chain-of-Skill and each edge corresponds to a skill applied between a pair of objects. Each node $v_i \in \mathcal{V}$ is encoded by a pre-trained RoBERTa [Liu *et al.*, 2019] model, resulting in an object embedding $\mathbf{x}_{v_i} \in \mathbb{R}^d$. Similarly, each skill phrase corresponding to edge r_{ij} is also encoded as a context-aware embedding $\mathbf{x}_{r_{ij}} \in \mathbb{R}^d$.

Motivated by the message propagation strategy in [Hamilton *et al.*, 2017], we enhance object and skill embeddings by aggregating contextual information across the graph. To incorporate the relational context of a skill edge r_{ij}, we update its representation based on the embeddings of its affiliated subject and object nodes:

$$\hat{\mathbf{x}}_{r_{ij}} = \mathbf{x}_{r_{ij}} + \mathcal{H}_r\left([\mathbf{x}_{v_i}; \mathbf{x}_{r_{ij}}; \mathbf{x}_{v_j}]\right), \tag{3}$$

where $\mathcal{H}_r$ is a learnable linear projection and $\hat{\mathbf{x}}_{r_{ij}} \in \mathbb{R}^d$ denotes the context-enhanced skill embedding. Subsequently, we update each object node v_i by attending to its neighboring nodes $\Psi(i)$ and their associated skill edges r_{ij} with a tailored attention mechanism, resulting in an updated context-aware object embedding:

$$\hat{\mathbf{x}}_{v_i} = \mathbf{x}_{v_i} + \sum_{j \in \Psi(i)} w_{v_{ij}} \mathcal{H}_v\left([\mathbf{x}_{v_j}; \hat{\mathbf{x}}_{r_{ij}}]\right), \tag{4}$$

where $\mathcal{H}_v$ is another projection layer and $w_{v_{ij}}$ is the attention weight computed as:

$$w_{v_{ij}} = \underset{j \in \Psi(i)}{\text{softmax}}\left(\mathcal{H}_v([\mathbf{x}_{v_i}; \hat{\mathbf{x}}_{r_{ij}}])^T \mathcal{H}_v([\mathbf{x}_{v_j}; \hat{\mathbf{x}}_{r_{ij}}])\right). \tag{5}$$

This message passing protocol enables each representation to be dynamically updated based on its semantic and procedural context, captured by the skill dependencies in the graph.

Chain-of-Skill Noise Watermark. Unlike unstable visual inputs that are susceptible to domain shifts, the noise space provides an inherently controllable and semantically neutral interface for incorporating high-level reasoning signals. Therefore, the reasoning-aware representation of the chain-of-skill structure is projected through a linear layer to obtain the chain-of-skill noise watermark δ_{lang}, which is injected into the noised trajectory tokens to guide temporally coherent, multi-step reasoning throughout the denoising process of the 3D Relative Denoising Transformer [Ke *et al.*, 2024]. Notably, since the chain-of-skill Noise Watermark δ_{lang} is learned during training, the entire chain-of-skill encoding pipeline (e.g., language reasoning graph construction and propagation) is no longer required at inference time. During deployment, the model simply reuses the learned watermark and injects it as a lightweight additive prior into the initial noise, enabling efficient inference without incurring additional computational overhead.

4.2 Self-evolving Policy Knowledge Bank

Motivation. Relying solely on the fixed Chain-of-Skill Noise Watermark is insufficient to provide adaptive, real-time guidance during long-horizon denoising, nor can it effectively correct drifting action trajectories. To this end, motivated by the unique capability of Vector Quantized Variational Autoencoder (VQ-VAE) [Van Den Oord *et al.*, 2017] to preserve essential data semantics in a compact latent space, we introduce a memory-augmented VQ-VAE that learns a discrete, self-evolving policy knowledge bank $\mathcal{C}$. This knowledge bank captures reusable and generalizable skill primitives distilled from historical experience, which are leveraged to guide the denoising process and to refine the denoised trajectory.

Encoder architecture. As shown in Fig. 2 (c), we take as visual input the current and all preceding observation images $I \in \mathbb{R}^{(t_I+1)\times h_I \times w_I \times c_I}$. Each observation is first partitioned into non-overlapping patches: the current frame yields patches of size $w_p \times h_p \times c_p$, and the preceding frames yield patches of size $t_p \times w_p \times h_p \times c_p$. Subsequently, every patch is flattened and linearly projected into a d_z-dimensional token. To explicitly disentangle spatial and temporal modeling, we reorganize the patch tokens into a tensor of shape $(t_z{+}1) \times (w_z h_z) \times d_z$, where $t_z = \frac{t_I}{t_p}$, $w_z = \frac{w_I}{w_p}$, and $h_z = \frac{h_I}{h_p}$ denote the numbers of temporal and spatial tokens after patch partitioning. On top of this representation, we apply Transformer layers in two sequential stages. First, a Spatial Transformer performs all-to-all self-attention over spatial tokens within each frame to capture intra-frame contextual and geometric relationships. Next, a Causal Transformer operates along the temporal dimension with a causal mask, modeling inter-frame dependencies such that each spatial token attends only to tokens from preceding frames. This hierarchical design enables the model to learn historical spatio-temporal representations $M \in \mathbb{R}^{(t_z+1)\times w_z h_z \times d_z}$ that encode spatial relations between actions and their environments, as well as causal dependencies across actions and temporal scene changes over extended horizons.

To achieve visuomotor alignment, we design a Visuomotor Transformer that leverages rich historical spatio-temporal

	Avg. S.R.↑	Avg. Rank↓	Push Buttons	Slide Block	Sweep to Dustpan	Meat off Grill	Turn Tap	Put in Drawer	Close Jar	Drag Stick
C2F-ARM-BC [James *et al.*, 2022]	20.1	9.0	72.0	16.0	0.0	20.0	68.0	4.0	24.0	24.0
PerAct [Shridhar *et al.*, 2023]	49.4	7.6	92.8$_{\pm 3.0}$	74.0$_{\pm 13.0}$	52.0$_{\pm 0.0}$	70.4$_{\pm 2.0}$	88.0$_{\pm 4.4}$	51.2$_{\pm 4.7}$	55.2$_{\pm 4.7}$	89.6$_{\pm 4.1}$
RVT [Goyal *et al.*, 2023]	62.9	6.3	**100.0**$_{\pm 0.0}$	81.6$_{\pm 5.4}$	72.0$_{\pm 0.0}$	88.0$_{\pm 2.5}$	93.6$_{\pm 4.1}$	88.0$_{\pm 5.7}$	52.0$_{\pm 2.5}$	99.2$_{\pm 1.6}$
Diffusion Policy [Goyal *et al.*, 2023]	64.4	6.1	97.2$_{\pm 3.0}$	74.2$_{\pm 4.3}$	80.3$_{\pm 1.0}$	90.8$_{\pm 2.7}$	88.4$_{\pm 3.9}$	95.2$_{\pm 4.9}$	67.0$_{\pm 3.5}$	100.0$_{\pm 0.0}$
3D-MVP [Qian *et al.*, 2024]	67.5	4.8	**100.0**	48.0	80.0	96.0	96.0	**100.0**	76.0	**100.0**
3D Diffuser Actor [Ke *et al.*, 2024]	81.3	3.4	98.4$_{\pm 2.0}$	97.6$_{\pm 3.2}$	84.0$_{\pm 4.4}$	96.8$_{\pm 1.6}$	99.2$_{\pm 1.6}$	96.0$_{\pm 3.6}$	96.0$_{\pm 2.5}$	**100.0**$_{\pm 0.0}$
RVT-2 [Goyal *et al.*, 2024]	81.4	3.3	**100.0**$_{\pm 0.0}$	92.0$_{\pm 2.8}$	**100.0**$_{\pm 0.0}$	99.0$_{\pm 1.7}$	99.0$_{\pm 1.7}$	96.0$_{\pm 2.0}$	**100.0**$_{\pm 0.0}$	99.0$_{\pm 1.7}$
DynaRend [Tian *et al.*, 2025]	83.2	2.5	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	93.6$_{\pm 5.4}$	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	99.2$_{\pm 1.8}$	91.2$_{\pm 4.4}$	**100.0**$_{\pm 0.0}$
ROITDP	**86.5**	**1.9**	**100.0**$_{\pm 0.0}$	98.6$_{\pm 1.2}$	95.3$_{\pm 2.8}$	**100.0**$_{\pm 0.0}$	99.6$_{\pm 1.2}$	**100.0**$_{\pm 0.0}$	94.3$_{\pm 3.1}$	**100.0**$_{\pm 0.0}$

	Put in Safe	Place Wine	Screw Bulb	Open Drawer	Stack Blocks	Stack Cups	Put in Cupboard	Insert Peg	Sort Shape	Place Cups
C2F-ARM-BC [James *et al.*, 2022]	12.0	8.0	8.0	20.0	0.0	0.0	0.0	4.0	8.0	0.0
PerAct [Shridhar *et al.*, 2023]	84.0$_{\pm 3.6}$	44.8$_{\pm 7.8}$	17.6$_{\pm 2.0}$	88.0$_{\pm 5.7}$	26.4$_{\pm 3.2}$	2.4$_{\pm 2.0}$	28.0$_{\pm 4.4}$	5.6$_{\pm 4.1}$	16.8$_{\pm 4.7}$	2.4$_{\pm 3.2}$
RVT [Goyal *et al.*, 2023]	91.2$_{\pm 3.0}$	91.0$_{\pm 5.2}$	48.0$_{\pm 5.7}$	71.2$_{\pm 6.9}$	28.8$_{\pm 3.9}$	26.4$_{\pm 8.2}$	49.6$_{\pm 3.2}$	11.2$_{\pm 3.0}$	36.0$_{\pm 2.5}$	4.0$_{\pm 2.5}$
Diffusion Policy [Goyal *et al.*, 2023]	91.6$_{\pm 4.4}$	89.0$_{\pm 6.3}$	62.8$_{\pm 4.6}$	63.2$_{\pm 5.1}$	20.6$_{\pm 3.2}$	31.4$_{\pm 6.2}$	56.6$_{\pm 4.5}$	20.2$_{\pm 2.4}$	25.4$_{\pm 4.7}$	5.6$_{\pm 3.2}$
3D-MVP [Qian *et al.*, 2024]	92.0	**100.0**	60.0	84.0	40.0	36.0	60.0	20.0	28.0	4.0
3D Diffuser Actor [Ke *et al.*, 2024]	97.6$_{\pm 2.0}$	93.6$_{\pm 4.8}$	82.4$_{\pm 2.0}$	89.6$_{\pm 4.1}$	68.3$_{\pm 3.3}$	47.2$_{\pm 8.5}$	85.6$_{\pm 4.1}$	**65.6**$_{\pm 4.1}$	44.0$_{\pm 4.4}$	24.0$_{\pm 7.6}$
RVT-2 [Goyal *et al.*, 2024]	96.0$_{\pm 2.8}$	95.0$_{\pm 3.3}$	88.0$_{\pm 4.9}$	74.0$_{\pm 11.8}$	80.0$_{\pm 2.8}$	69.0$_{\pm 5.9}$	66.0$_{\pm 4.5}$	40.0$_{\pm 0.0}$	35.0$_{\pm 7.1}$	**38.0**$_{\pm 4.5}$
DynaRend [Tian *et al.*, 2025]	96.0$_{\pm 4.0}$	98.4$_{\pm 2.2}$	88.0$_{\pm 6.3}$	89.6$_{\pm 2.2}$	71.2$_{\pm 3.3}$	82.4$_{\pm 3.6}$	**87.2**$_{\pm 1.8}$	31.2$_{\pm 3.3}$	44.8$_{\pm 3.3}$	25.6$_{\pm 5.4}$
ROITDP	**99.3**$_{\pm 3.2}$	93.6$_{\pm 2.4}$	**95.2**$_{\pm 3.3}$	**92.2**$_{\pm 5.0}$	**83.0**$_{\pm 2.6}$	**90.8**$_{\pm 3.1}$	**87.2**$_{\pm 2.1}$	43.8$_{\pm 1.9}$	**52.3**$_{\pm 2.3}$	31.8$_{\pm 4.7}$

Table 1: Evaluation results on 18 RLBench tasks. Each task is evaluated with 25 rollouts under 5 different seeds. We report the average success rate and standard deviation for all tasks.

context to refine the ongoing denoised trajectory. Implemented as a stack of self- and cross-attention layers, it enables explicit and fine-grained interaction between visual perception and action streams. Through this design, the Visuomotor Transformer learns to reason over task-relevant visual relational cues and refine the action representation, producing a unified latent embedding $z_e(\hat{A}_0) \in \mathbb{R}^w$. This continuous latent embedding is subsequently quantized to capture reusable policy knowledge by selecting the nearest entry c_u from the codebook $\mathcal{C} = \{\mathbf{c}_u \in \mathbb{R}^w, u = 1, \ldots, U\}$ with vocabulary size U, according to:

$$z_q(\hat{A}_0) = c_u, \quad \text{where } u = \arg\min_j \|z_e(\hat{A}_0) - c_j\|_2. \tag{6}$$

The resulting code vectors form a discrete policy representation space, i.e., a self-evolving policy knowledge bank, where each code corresponds to a disentangled skill primitive. These skill primitives capture frequently occurring and generalizable action patterns distilled from historical trajectories (e.g., grasp, place, and push) and serve as a compact memory for guiding and refining action generation.

Decoder architecture. After quantization, we apply the NSVQ technique [Vali and Bäckström, 2022] to prevent gradient collapse and encourage diverse codebook utilization, which perturbs the encoder embedding with a normalized Gaussian noise vector:

$$\tilde{z}_e(\hat{A}_0) = z_e(\hat{A}_0) + \frac{\|z_e(\hat{A}_0) - z_q(\hat{A}_0)\|_2}{\|\xi\|_2}\,\xi, \tag{7}$$

where $\xi \sim \mathcal{N}(0, I)$. The decoder subsequently reconstructs the action trajectory via:

$$\tilde{A}_0 = D\big(\mathrm{Attn}(\mathrm{sg}[M], \tilde{z}_e(\hat{A}_0), \tilde{z}_e(\hat{A}_0))\big), \tag{8}$$

where stop gradient (sg) is applied to M to avoid representation collapse, and cross-attention is used to attend $\tilde{z}_e(\hat{A}_0)$ conditioned on M. Unlike the encoder, the decoder D only includes a Visuomotor Transformer, which reconstructs the denoised trajectory $\tilde{A}_0$ in an auto-regressive manner over time. The loss function is as follows:

$$\mathcal{L}_{\text{VQ}} = \big\|\hat{A}_0 - \tilde{A}_0\big\|_2^2 + \big\|\mathrm{sg}\big[\tilde{z}_e(\hat{A}_0)\big] - c_u\big\|_2^2 + \big\|\tilde{z}_e(\hat{A}_0) - \mathrm{sg}\big[c_u\big]\big\|_2^2. \tag{9}$$

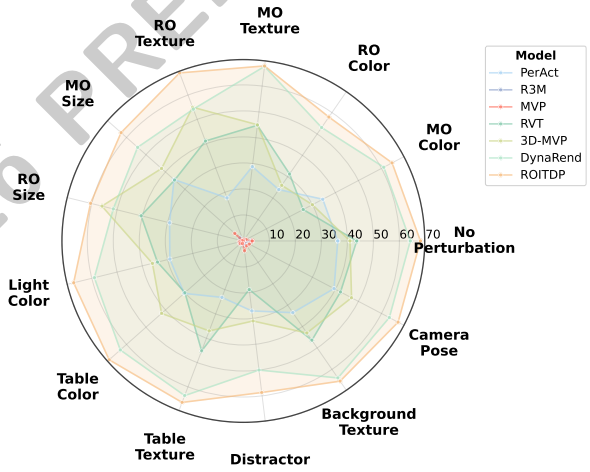

Figure 3: Results on COLOSSEUM.

4.3 Training Objective

Following the standard diffusion policy training objective (Eq. (1)), we apply the same noise estimation loss to both the denoised trajectory produced by the Denoising Transformer and the refined trajectory generated by the VQ-VAE branch, resulting in two diffusion losses, $\mathcal{L}_{\text{DP}}$ and $\mathcal{L}_{\text{DP-VQ}}$. The final training objective is the sum of three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DP}} + \mathcal{L}_{\text{DP-VQ}} + \mathcal{L}_{\text{VQ}}. \tag{10}$$

Modules				RLBench	COLOSSEUM	
Row	COS-NW	SPKB	Strategy	Avg. S.R.(%)	Avg. S.R.(%)	Δ
1	✗	✗	-	81.3	59.7	−9.5
2	✓	✗	Image Watermark	81.9	62.0	−8.9
3	✓	✗	Noise Watermark	84.2	65.0	−5.8
4	✗	✓	Only SPKB Correction	81.8	61.6	−9.0
5	✗	✓	Only SPKB Guidance	82.7	62.5	−8.7
6	✗	✓	Full SPKB	83.6	63.8	−6.9
7	✓	✓	-	86.5	68.1	−3.4

Table 2: Ablation study of the proposed modules on RLBench and Colosseum. We report the average success rate (Avg. S.R.) on RLBench, the average success rate under the *No Perturbation* setting on Colosseum, as well as the average performance degradation (Δ) under environmental perturbations.

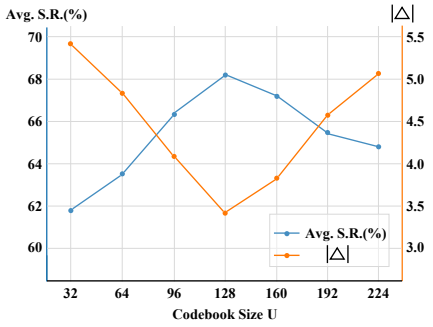

Figure 4: Specificity–Generalizability trade-off with varying codebook size U in the policy knowledge bank.

5 Experiments

5.1 Simulation Experiments

Datasets and Evaluation Metrics. We conduct all simulation experiments on two challenging robotic manipulation benchmarks: **RLBench** [James *et al.*, 2020] and **COLOSSEUM** [Pumacay *et al.*, 2024]. RLBench is a large-scale simulated manipulation suite covering tasks such as grasping, pushing, opening, and placing. Following prior work [Goyal *et al.*, 2023], we adopt an 18-task subset and collect 100 expert demonstrations per task for training. Each policy is evaluated over 25 rollout episodes, and success is determined by task-specific goal satisfaction. COLOSSEUM provides a complementary benchmark for testing policy generalization under visual and physical perturbations. It includes 20 object-centric manipulation tasks evaluated under 12 types of environment variations in color, texture, size, and lighting.

Implementation Details. We implement the proposed ROITDP model using PyTorch and train it end-to-end on 4 NVIDIA A6000 GPUs. The model is optimized using the AdamW optimizer with a batch size of 240 and a learning rate of 1e-4 with cosine decay scheduling. Each image is rendered at a resolution of 256×256. Following prior work [Ke *et al.*, 2024], all simulation experiments are built upon the CoppeliaSim physics simulator and employ a 7-DoF *Franka Emika Panda* robot equipped with a parallel gripper, operating in a fixed tabletop workspace.

Results on RLBench. The experimental results on RLBench are shown in Tab. 1, with the following key observations: (i) Our method achieves state-of-the-art performance with an average success rate of 86.5%, surpassing the best baseline DynaRend [Tian *et al.*, 2025] by 3.3%. (ii) On tasks requiring precise coordination between spatial reasoning and action sequencing (e.g., *Screw Bulb*), ROITDP attains a success rate of 95.2%, outperforming DynaRend and demonstrating that the chain-of-skill noise watermark effectively enhances the model's multi-step reasoning capability. (iii) On more challenging tasks that involve long-horizon reasoning or complex object interactions (e.g., *Stack Cups*), our method consistently outperforms current state-of-the-art approaches by a notable margin. This improvement suggests that the proposed self-evolving policy knowledge bank provides strong historical guidance during the denoising process, while enabling further refinement and correction of the denoised action trajectories.

Results on COLOSSEUM. Fig. 3 presents the success rates of different methods on the COLOSSEUM dataset. ROITDP consistently outperforms both 2D pretraining methods (e.g., MVP [Xiao *et al.*, 2022], R3M [Nair *et al.*, 2022]) and 3D baselines (e.g., 3D-MVP [Qian *et al.*, 2024] and DynaRend [Tian *et al.*, 2025]) across most categories of environmental perturbations. Unlike previous methods [Qian *et al.*, 2024; Tian *et al.*, 2025] that exhibit noticeable performance fluctuations, our ROITDP maintains consistently high success rates, validating its robust generalization capability under diverse domain shifts. We attribute these gains to the proposed chain-of-skill noise watermark and self-evolving policy knowledge bank, which provide strong historical guidance during the denoising process and enable more stable action generation in challenging environments.

5.2 Ablation Studies

To validate the effectiveness of each proposed module within our ROITDP framework, we conduct ablation studies on the RLBench and Colosseum datasets, as shown in Tab. 2.

Effectiveness of Chain-of-Skill Noise Watermark (COS-NW). Comparing Rows 1–3, we observe that introducing the skill-chain watermark consistently improves performance over the vanilla diffusion policy. Moreover, noise-level watermarking significantly mitigates performance degradation under environmental perturbations. This demonstrates that skill guidance injected into the noise space is more robust to visual-space perturbations, while enabling the policy to learn temporally structured skill chains that directly guide multi-step action reasoning during the denoising process.

Effectiveness of Self-evolving Policy Knowledge Bank (SPKB). The impact of SPKB is evident when comparing Rows 4–6. We report the performance under three configurations: (i) SPKB with VQ-VAE-based correction only; (ii) SPKB with policy knowledge bank guidance during denoising only; and (iii) the full SPKB model that combines both correction and guidance. Specifically, SPKB guidance (Row 5) achieves higher gains than SPKB correction (Row 4), indicating that leveraging learned skill primitives to guide the denoising process is more effective than post-hoc correction alone. Additionally, combining SPKB correction and guidance (Full SPKB, Row 6) further improves performance, yielding higher success rates on both RLBench and Colosseum while mitigating performance degradation under per-

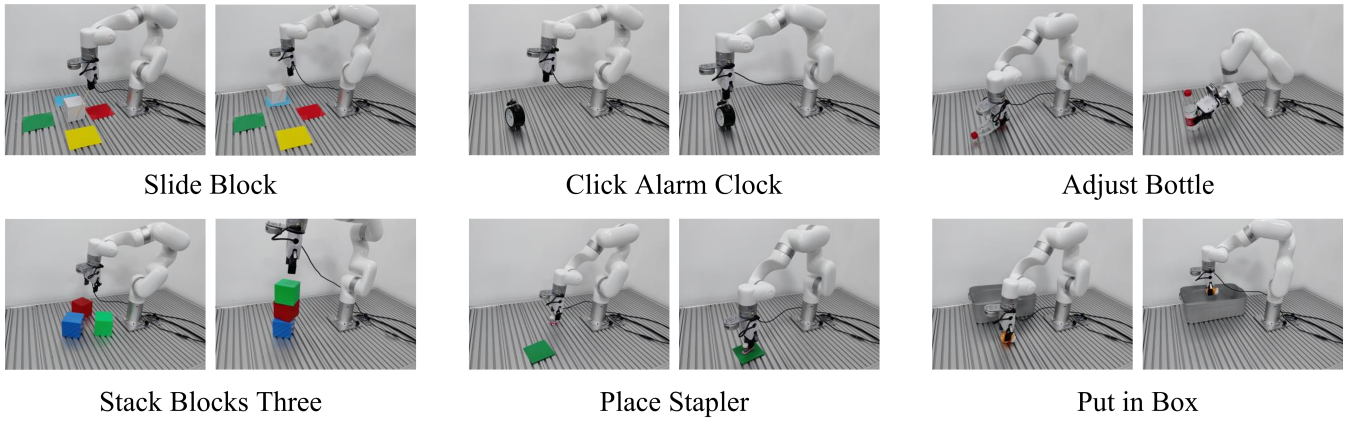

Figure 5: Robot setup and examples. We evaluate on six manipulation tasks: Slide Block, Click Alarm Clock, Adjust Bottle, Stack Blocks Three, Place Stapler, and Put in Box.

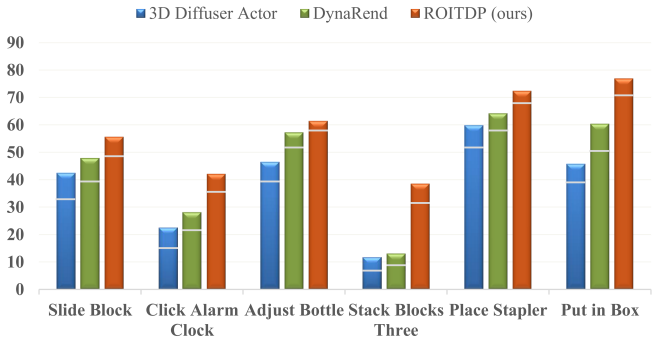

Figure 6: Real-world experiment performance. We report the average success rates over 20 rollouts for each task. The portion above the gray line indicates performance degradation under out-of-distribution conditions.

turbations. These results suggest that the self-evolving policy knowledge bank not only provides effective memory of reusable skill primitives, but also enables iterative refinement and self-correction of action trajectories, leading to more robust and generalizable policy learning.

Specificity–Generalizability Trade-off. Fig. 4 illustrates the impact of varying the codebook size U in the policy knowledge bank on both the average task success rate and the absolute average performance degradation under distribution shifts. Interestingly, the vocabulary size U can be viewed as a controllable information bottleneck that modulates the granularity of policy representations. Larger codebooks become overly expressive, retaining nearly all low-level variations in motion, which may lead to overfitting or redundancy. Smaller codebooks promote compact and abstract representations by quantizing continuous trajectories into reusable skill primitives (e.g., *grasp*, *place*, *push*), while filtering out minor trajectory variations and retaining semantically meaningful skills. When U is too small, however, the codebook may collapse into overly generic tokens, thus limiting the model's capacity to capture nuanced behaviors.

5.3 Real-world Experiments

Environmental Setup. As illustrated in Fig. 5 and Fig. 6, we evaluate ROITDP on six challenging real-world manipulation tasks and compare it with representative state-of-the-art baselines. All experiments are conducted on an xArm6 robot equipped with a parallel-jaw gripper in a fixed tabletop setting. The robot is equipped with two fixed external ZED cameras for multi-view perception and an Intel RealSense L515 wrist-mounted depth camera for close-range sensing, all calibrated and synchronized in a shared coordinate frame. For each task, we collect 50 expert demonstrations with randomized object placements and initial configurations.

Quantitative Results. ROITDP consistently achieves higher success rates with reduced performance degradation under out-of-distribution conditions across all evaluated tasks compared to prior methods in Fig. 6. The gains are particularly pronounced in long-horizon tasks with complex skill compositions, such as *Stack Blocks Three* and *Put in Box*. These improvements stem from the proposed Chain-of-Skill Noise Watermark, which provides temporally coherent reasoning guidance, and the Self-evolving Policy Knowledge Bank, which enables adaptive reuse of learned skill primitives. Together, they allow ROITDP to generate more stable and robust behaviors in real-world settings.

6 Conclusion

In this paper, we present ROITDP, a reasoning- and memory-augmented diffusion policy framework for generalizable robotic manipulation. By injecting skill-level reasoning into the diffusion noise space and maintaining an adaptive policy knowledge bank, ROITDP enables diffusion policies to "raise one and infer three" from limited demonstrations. Extensive experiments demonstrate that our approach achieves strong generalization, particularly under long-horizon tasks and distribution shifts. We believe that explicitly integrating reasoning and memory provides a promising direction for advancing diffusion-based policy learning toward more robust, transferable, and scalable robotic behaviors.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600) and the National Natural Science Foundation of China (U25B2048, 62132016).

References

[Brohan *et al.*, 2022] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[Cheng *et al.*, 2024] Jie Cheng, Yingbing Chen, and Qifeng Chen. Pluto: Pushing the limit of imitation learning-based planning for autonomous driving. *arXiv preprint arXiv:2404.14327*, 2024.

[Chi *et al.*, 2023] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[Goyal *et al.*, 2023] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.

[Goyal *et al.*, 2024] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024.

[Haldar *et al.*, 2023] Siddhant Haldar, Jyothish Pari, Anant Rai, and Lerrel Pinto. Teach a robot to fish: Versatile imitation from one minute of demonstrations. *arXiv preprint arXiv:2303.01497*, 2023.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[Hu *et al.*, 2023] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.

[James *et al.*, 2020] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

[James *et al.*, 2022] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.

[Jiang *et al.*, 2025] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Jim Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16923–16930. IEEE, 2025.

[Ke *et al.*, 2024] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

[Li *et al.*, 2023] Tingguang Li, Yizheng Zhang, Chong Zhang, Qingxu Zhu, Jiapeng Sheng, Wanchao Chi, Cheng Zhou, and Lei Han. Learning terrain-adaptive locomotion with agile behaviors by imitating animals. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 339–345. IEEE, 2023.

[Li *et al.*, 2024] Zhuoling Li, Liangliang Ren, Jinrong Yang, Yong Zhao, Xiaoyang Wu, Zhenhua Xu, Xiang Bai, and Hengshuang Zhao. Vip: Vision instructed pre-training for robotic manipulation. *arXiv preprint arXiv:2410.07169*, 2024.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Lyu *et al.*, 2025] Guangtao Lyu, Chenghao Xu, Qi Liu, Jiexi Yan, Muli Yang, Fen Fang, and Cheng Deng. Tempo as the stable cue: Hierarchical mixture of tempo and beat experts for music to 3d dance generation. *arXiv preprint arXiv:2512.18804*, 2025.

[Nair *et al.*, 2022] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[Pumacay *et al.*, 2024] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

[Qian *et al.*, 2024] Shengyi Qian, Kaichun Mo, Valts Blukis, David F Fouhey, Dieter Fox, and Ankit Goyal. 3d-mvp: 3d multiview pretraining for robotic manipulation. *arXiv preprint arXiv:2406.18158*, 2024.

[Schuster *et al.*, 2015] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[Shridhar *et al.*, 2023] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[Tang *et al.*, 2024] Annan Tang, Takuma Hiraoka, Naoki Hiraoka, Fan Shi, Kento Kawaharazuka, Kunio Kojima, Kei Okada, and Masayuki Inaba. Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13107–13114. IEEE, 2024.

[Tian *et al.*, 2025] Jingyi Tian, Le Wang, Sanping Zhou, Sen Wang, Jiayi Li, and Gang Hua. Dynarend: Learning 3d dynamics via masked future rendering for robotic manipulation. *arXiv preprint arXiv:2510.24261*, 2025.

[Vali and Bäckström, 2022] Mohammad Hassan Vali and Tom Bäckström. Nsvq: Noise substitution in vector quantization for machine learning. *IEEE Access*, 10:13598–13610, 2022.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2024] Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kelestemur, Jiuguang Wang, and Yunzhu Li. Gendp: 3d semantic fields for category-level generalizable diffusion policy. In *8th Annual Conference on Robot Learning*, volume 2, 2024.

[Wang *et al.*, 2025] Yuxuan Wang, Aming Wu, Muli Yang, Yukuan Min, Yihang Zhu, and Cheng Deng. Reasoning mamba: Hypergraph-guided region relation calculating for weakly supervised affordance grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27618–27627, 2025.

[Xiao *et al.*, 2022] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[Xu *et al.*, 2024] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024.

[Yan *et al.*, 2025] Ge Yan, Yueh-Hua Wu, and Xiaolong Wang. Dnact: Diffusion guided multi-task 3d policy learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9464–9471. IEEE, 2025.

[Yang *et al.*, 2024] Jingyun Yang, Zi-ang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning. *arXiv preprint arXiv:2407.01479*, 2024.

[Ze *et al.*, 2024] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

[Ze *et al.*, 2025] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3d diffusion policies. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2873–2880. IEEE, 2025.

[Zhang *et al.*, 2024] Zhilong Zhang, Ruifeng Chen, Junyin Ye, Yihao Sun, Pengyuan Wang, Jingcheng Pang, Kaiyuan Li, Tianshuo Liu, Haoxin Lin, Yang Yu, et al. Whale: Towards generalizable and scalable world models for embodied decision-making. *arXiv preprint arXiv:2411.05619*, 2024.

[Zhu *et al.*, 2025a] Xupeng Zhu, Fan Wang, Robin Walters, and Jane Shi. Se (3)-equivariant diffusion policy in spherical fourier space. *arXiv preprint arXiv:2507.01723*, 2025.

[Zhu *et al.*, 2025b] Yihang Zhu, Jinhao Zhang, Yuxuan Wang, Aming Wu, and Cheng Deng. Vgmamba: Attribute-to-location clue reasoning for quantity-agnostic 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5295–5304, October 2025.