2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SOE: Sample-Efficient Robot Policy Self-Improvement via On-Manifold Exploration

Yang Jin[1], Jun Lv[1,3], Han Xue[1], Wendi Chen[1,2], Chuan Wen[1†], Cewu Lu[1,2,3†]
[1]Shanghai Jiao Tong University [2]Shanghai Innovation Institute [3]Noematrix Ltd. [†]Corresponding Author
ericjin2002.github.io/SOE

***Abstract*— Intelligent agents progress by continually refining their capabilities through actively exploring environments. Yet robot policies often lack sufficient exploration capability due to action mode collapse. Existing methods that encourage exploration typically rely on random perturbations, which are unsafe and induce unstable, erratic behaviors, thereby limiting their effectiveness. We propose Self-Improvement via On-Manifold Exploration (SOE), a framework that enhances policy exploration and improvement in robotic manipulation. SOE learns a compact latent representation of task-relevant factors and constrains exploration to the manifold of valid actions, ensuring safety, diversity, and effectiveness. It can be seamlessly integrated with arbitrary policy models as a plug-in module, augmenting exploration without degrading the base policy performance. Moreover, the structured latent space enables human-guided exploration, further improving efficiency and controllability. Extensive experiments in both simulation and real-world tasks demonstrate that SOE consistently outperforms prior methods, achieving higher task success rates, smoother and safer exploration, and superior sample efficiency. These results establish on-manifold exploration as a principled approach to sample-efficient policy self-improvement.**

I. Introduction

"We want AI agents that can discover like we can, not which contain what we have discovered."
— Richard Sutton, The Bitter Lesson

In recent years, data-driven robot learning [10, 53, 5, 7, 25] has attracted considerable attention, particularly for its potential to enhance robotic manipulation capabilities through large-scale data collection and training. By modeling visuomotor behaviors with neural networks, these approaches allow robot policies to learn from expert demonstrations and achieve near-human performance across a variety of tasks.

Despite these advances, most existing methods still rely heavily on human teleoperation for data acquisition [53, 13] and policy refinement [30, 31], which presents several challenges. A primary concern is the high cost of teleoperation, as it typically requires skilled operators and specialized equipment, thereby limiting the scalability of data collection. More critically, teleoperated demonstrations often fail to cover the diverse scenarios a robot could encounter in the real world, resulting in distributional bias [52] and compounding error [39]. The problem is further exacerbated by the fact that human operators may act based on contextual cues inaccessible to robot sensors. Robots, on the other hand, may internalize human habits rather than task-relevant behaviors. As a result, simply scaling up teleoperated data is not the optimal path toward improving policy performance.

Instead of passively imitating human-provided behavior, a line of research addresses this challenge by enabling robot policy self-improvement [6, 23, 35, 32]—actively exploring the environment to collect diverse experience and leveraging that experience to refine policies. Under this paradigm, robots can autonomously discover novel behaviors that go beyond the coverage of human demonstrations. By iteratively practicing the learned behaviors, they also develop a deeper understanding of the natural variability in their actions, ultimately leading to a more robust and resilient policy.

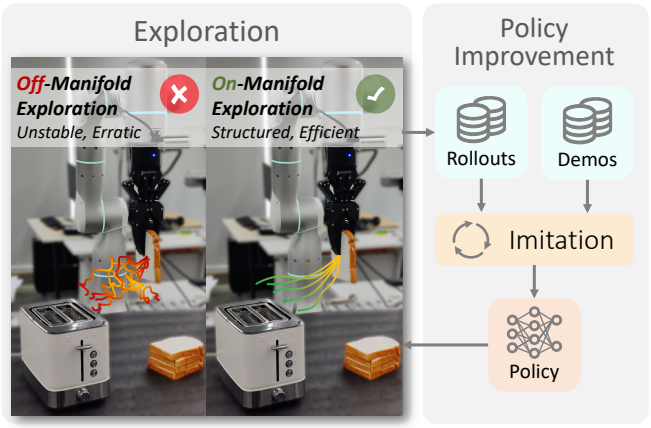


Fig. 1: **Overview of SOE.** By constraining exploration to the manifold of valid actions, our approach generates diverse yet temporally coherent behaviors, enabling structured and efficient exploration. The collected rollout data is used to refine the policy, leading to efficient self-improvement.

The key to sample-efficient robot policy self-improvement lies in effective exploration. Prior work [3, 23] has shown that imitation-learned policies often overfit demonstrations, collapse into single-modal motions, and fail to produce diverse behaviors. Without proper exploration, these policies tend to repeat failed behaviors, limiting their ability to discover improved solutions. While random exploration strategies can occasionally yield novel behaviors [29], they are generally ineffective in high-dimensional action spaces [28] and can pose safety risks in real-world deployment [16], causing potential hardware damage. This necessitates a more structured approach to exploration—one that ensures safety and effectiveness without sacrificing the diversity of experiences.

To this end, we propose *SOE*, a novel framework for *Sample-Efficient Robot Policy **S**elf-improvement via **O**n-Manifold **E**xploration*. The core idea of our method is to ensure that exploration remains constrained to the manifold of valid actions—critical for both safety and effectiveness.

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

Prior works often perturb the action space directly [29] or inject random noise [23], leading to temporally inconsistent and unsafe behaviors, particularly under "action chunking" representations [53]. In contrast, we perform exploration in a compact latent space learned through a variational information bottleneck (VIB). The latent representation in this space preserves only task-essential information in observation while discarding irrelevant details, ensuring exploration remains structured and efficient. As illustrated in Fig. 1, by operating on this latent representation, our framework enables effective on-manifold exploration and more robust policy improvement. Furthermore, we demonstrate that in the latent space, action chunks are naturally disentangled into distinct modes. Leveraging this property, we achieve controllable exploration, which allows users to guide exploration toward preferred directions, thereby enhancing interpretability and further boosting sample efficiency. Implemented as a plug-in module, our approach can be seamlessly integrated with existing imitation learning algorithms and jointly optimized, without any degradation in their performance.

To evaluate the effectiveness of our method, we conduct extensive experiments across a variety of robot manipulation tasks in both simulation and real world. The results show that *SOE* consistently outperforms prior exploration methods in effectiveness, motion smoothness, and sample efficiency. With just one round of policy self-improvement, our method achieves substantial gains over the base policy, including an average relative improvement of 50.8% on real-world tasks. Additional experiments in simulation and ablation studies further confirm multi-round performance improvements and the contribution of each component in our framework. Collectively, these findings demonstrate that on-manifold exploration provides a structured, safe, and effective approach to sample-efficient robot policy self-improvement.

II. Related Work

A. Imitation Learning for Robot Manipulation

Imitation learning is an extensively studied approach for training robot policies by mimicking expert demonstrations. Early methods, such as behavior cloning, directly learn a mapping from observations to actions using supervised learning. More recent approaches leverage advanced architectures, including transformers [53, 41, 7, 54, 18] and flow-based generative models [22, 10, 1, 51, 48], to capture multimodal visuomotor behaviors. These methods have demonstrated impressive performance on diverse manipulation tasks, promising a data-driven path for general-purpose robot learning [25, 5, 4, 46]. However, most of them still rely heavily on costly expert demonstrations, suffering from limited or biased data coverage, motivating the need for policy post-training.

B. Real-World Post-training of Robot Policy

Reinforcement learning (RL) [44, 17, 40, 29] is a widely used paradigm for post-training robot policies, enabling robots to learn from trial and error. However, RL typically requires a prohibitive amount of interactions to achieve satisfactory performance, limiting their practicality for real-world applications. To address this, researchers have explored sim-to-real transfer [3, 33], offline RL [14, 27, 26], residual policy learning [49, 19], and human-in-the-loop approaches [31, 9] to enhance sample efficiency. While effective in some cases, these methods face issues like extrapolation errors [26], safety concerns [16], and reliance on hand-crafted reward designs or labor-intensive human teleoperation. An alternative is to enable robot self-improvement through autonomous imitation learning [35, 6, 23, 30]. However, imitation-learned policies often lack behavioral diversity, highlighting the need for effective exploration.

C. Exploration in Robot Learning

Exploration is crucial yet challenging in robot learning, particularly in high-dimensional spaces. Most existing exploration approaches can be categorized into two lines: reward-based methods and sampling-based methods. The first line of methods [37, 12, 8, 38, 24, 15] achieves exploration by introducing some intrinsic rewards, based on novelty or curiosity, to incentivize the agent to visit unfamiliar or unpredictable states. The second line of methods diversifies action generation through strategies like epsilon-greedy [36, 47], Boltzmann exploration [43, 45], and goal-directed exploration [21], which are typically achieved by perturbing the action space or injecting noise into the policy. Despite the success of both lines of methods in various domains, they often struggle in real-world manipulation with continuous, chunked actions. Perhaps the most relevant work to ours is *SIME* [23], which perturbs diffusion policy conditions to induce modal-level exploration. Our method extends this idea by constraining exploration to the task manifold, enabling safer, more effective, and controllable exploration behaviors.

III. Preliminaries and Problem Formulation

In this section, we introduce the background of imitation learning and policy self-improvement.

Imitation Learning. Consider a robot manipulation task $\mathcal{T}$ along with a set of expert demonstrations $\mathcal{D}^e = \{\tau_i\}_{i=1}^N$, where each trajectory $\tau_i = (o_1, a_1, \ldots, o_T, a_T)$ consists of a sequence of observations $o_t \in \mathcal{O}$ and actions $a_t \in \mathcal{A}$. The goal of imitation learning is to learn a policy $\pi_\theta : \mathcal{O} \to \mathcal{A}$ that maps observations to actions by mimicking the expert behavior. This is typically achieved through optimizing the policy parameters θ by maximizing the likelihood of the expert actions given the corresponding observations:

$$\theta^* = \arg\max_\theta \mathbb{E}_{(o_t,a_t)\sim\mathcal{D}^e} \left[\log \pi_\theta(a_t|o_t)\right]. \tag{1}$$

Policy Self-Improvement. Beyond imitation learning, policy self-improvement aims to enhance the learned policy by enabling the robot to actively explore its environment and collect additional experience. By learning from those experiences, the robot is expected to improve its performance beyond the limitations of the initial expert demonstrations. This procedure also aligns with the rejection sampling fine-tuning (RFT) [50] paradigm in the large language model post-training literature, which has been proven effective in boosting model performance.

Formally, given the expert demonstration dataset $\mathcal{D}^e$, an imitation-learned policy π_0, potentially augmented with exploration mechanisms, interacts with the environment to collect a set of additional trajectories $\mathcal{D}^b$, which is merged with the original dataset to form an aggregated dataset $\mathcal{D} = \mathcal{D}^e \cup \mathcal{D}^b$. The policy is then refined on this aggregated dataset to obtain an improved policy π_1. This process can be iteratively repeated, yielding a sequence of progressively enhanced policies: $\pi_0, \pi_1, \pi_2, \ldots, \pi_m$.

The objective of policy self-improvement is to maximize the success rate of the policy π_m while minimizing the number of interactions $|\mathcal{D}^b|$ required to achieve this improvement and ensuring the safety of the exploration process.

IV. Method

In this paper, we present *SOE*, a novel framework for sample-efficient robot policy self-improvement via on-manifold exploration. We begin by outlining the challenges of exploration in high-dimensional action spaces and introduce on-manifold exploration, which involves learning a compact latent representation that captures task-essential factors. Next, we describe how our exploration mechanism can be seamlessly integrated into existing imitation learning pipelines as a plug-in module, without degrading base policy performance. Finally, we introduce user-guided steering, which leverages the disentangled nature of the learned representation for controllable exploration.

A. On-Manifold Exploration

Exploration is critical for policy self-improvement, as it enables the robot to discover novel behaviors beyond the coverage of expert demonstrations. However, achieving effective exploration in action spaces is notoriously challenging, particularly when actions are continuous and temporally grouped in chunks. This is because the action space is typically vast, and the task-relevant manifold only occupies a small fraction of it. Directly injecting random noise into actions often drives the policy off the task manifold, leading to unsafe, jerky motions and ineffective exploration, as illustrated in Fig. 1. Perturbing latent observation embeddings may appear to be an alternative, but since these embeddings are usually entangled with redundant or correlated features, such perturbations can still yield unrealistic or out-of-distribution actions, as demonstrated in our experiments.

This motivates the need for learning a more structured, well-shaped feature space, which can effectively capture the underlying task manifold and facilitate sample-efficient on-manifold exploration. To address this, we propose learning a compact latent representation of observations that retains only the information essential for policy execution while discarding irrelevant, high-frequency details. This yields a more structured and robust target for perturbations, enabling effective on-manifold exploration. Formally, given observation O and action A, we define the latent variable Z via an encoder $p_\theta(z|o)$, optimized with the following objective:

$$\max_\theta I(Z;A) - \beta I(Z;O), \tag{2}$$

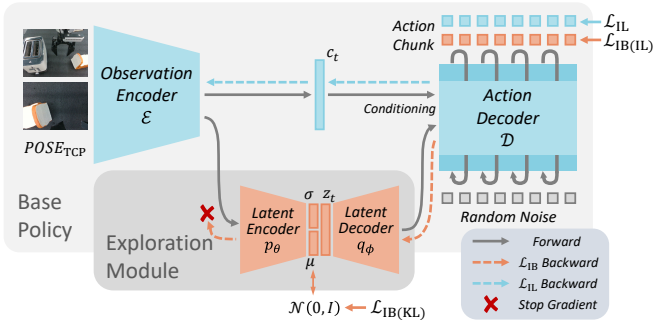


Fig. 2: **Dual-Path Architecture of *SOE*.** The observation embedding is processed through two parallel paths: the base path (top), responsible for stable policy execution, and the exploration path (bottom), responsible for generating diverse actions. Each path outputs distinct noise predictions and is optimized with a separate loss function.

where $I(\cdot;\cdot)$ denotes mutual information and β controls the trade-off between informativeness and compactness. This encourages Z to preserve action-relevant features while minimizing dependence on extraneous observation details. Following the variational information bottleneck (VIB) [2], we derive a tractable variational upper bound from Eqn. 2:

$$\mathcal{L}_{\text{IB}}(\theta,\phi) = \mathbb{E}_{(o,a)\sim\mathcal{D}}[-\mathbb{E}_{z\sim p_\theta(z|o)} \log q_\phi(a|z) + \beta \text{KL}[p_\theta(Z|o)||r(Z)]], \tag{3}$$

where $q_\phi(a|z)$ is a decoder that reconstructs actions from Z, and $r(Z)$ is a predefined prior. The first term, denoted as $\mathcal{L}_{\text{IB(IL)}}$, corresponds to the imitation loss in Eqn. 1, encouraging Z to retain action-relevant information. The second term, denoted as $\mathcal{L}_{\text{IB(KL)}}$, acts as a regularizer that penalizes deviations from the prior. In our implementation, the encoder p_θ is parameterized as a diagonal Gaussian model, and the prior $r(Z)$ is chosen as a d-dimensional isotropic Gaussian $\mathcal{N}(Z;0,I)$.

With the latent representation Z, the underlying low-dimensional task manifold is explicitly modeled. By sampling from p_θ and decoding through q_ϕ, we can generate diverse, on-manifold action proposals:

$$\mu_t, \sigma_t \sim p_\theta(Z|O=o_t), \quad z_t \sim \mathcal{N}(\mu_t, (\alpha\sigma_t)^2),$$
$$a_t \sim q_\phi(A|Z=z_t),$$

where μ_t and σ_t establish the local geometry of the task manifold around the current observation o_t, and α is a hyperparameter that controls the exploration scale. A larger α expands exploration to a broader neighborhood on the manifold, while a smaller α restricts exploration to local variations. By adjusting α, the exploration strategy can smoothly transition between conservative and aggressive, providing a flexible mechanism to balance safety and diversity.

Compared to naïve random noise injection, our approach leverages the structured distribution of the learned latent space to perform exploration via sampling. This ensures that the generated action proposals remain on the task manifold, avoiding unrealistic or unsafe behaviors, while still

supporting diverse and informative exploration, as illustrated in Fig. 1. With this on-manifold exploration mechanism, the robot can efficiently discover successful behaviors, enabling sample-efficient robot policy self-improvement.

B. Exploration as a Plug-in

To enhance the general applicability of the proposed approach, a central question is how to incorporate the exploration mechanism into standard imitation learning pipelines. Two challenges are particularly salient. First, the integration must preserve the expressive capacity of base policy, ensuring exploration does not compromise its fundamental performance. Second, the introduced component should impose minimal computational and optimization overhead, ideally allowing joint, end-to-end training with base policy.

We address these challenges by designing our exploration mechanism as a plug-in module that can be seamlessly integrated and jointly optimized with existing policy networks. As illustrated in Fig. 2, this is achieved through a dual-path architecture that supports both stable policy execution and diverse exploration. In this setup, the latent encoder p_θ and decoder q_ϕ form an auxiliary bypass alongside the original observation encoder $\mathcal{E}$ and action decoder $\mathcal{D}$ of the base policy, resulting in two parallel action-generation paths:

- **Base path**, where the embedding from the observation encoder $\mathcal{E}$ directly conditions the action decoder $\mathcal{D}$:

$$c_t = \mathcal{E}(o_t), \quad a_{t:t+H} = \mathcal{D}(c_t),$$

  where c_t is the observation embedding at time t and $a_{t:t+H}$ is the generated action chunk spanning H steps.
- **Exploration path**, where the observation embedding c_t is encoded into the latent space, perturbed stochastically, and then decoded back into a modified embedding $\tilde{c}_t$ for diverse action proposal generation:

$$c_t = \mathcal{E}(o_t), \quad \mu_t, \sigma_t = p_\theta(c_t), \quad z_t \sim \mathcal{N}(\mu_t, (\alpha\sigma_t)^2),$$
$$\tilde{c}_t = q_\phi(z_t), \quad \tilde{a}_{t:t+H} = \mathcal{D}(\tilde{c}_t).$$

The base path ensures that the core policy remains intact, while the exploration path facilitates the generation of varied action proposals through structured perturbations in the latent space. Together, they form a switchable exploration mechanism, allowing users to alternate between standard execution and active exploration as needed.

During training, both paths are optimized jointly. The base path is trained with a standard imitation loss $\mathcal{L}_{\text{IL}}$ (Eqn. 1), while the exploration path is trained with a variational information bottleneck loss $\mathcal{L}_{\text{IB}}$ (Eqn. 3), which combines action reconstruction with KL regularization. We instantiate the base policy as a diffusion policy [10], which leads to the following losses:

$$\mathcal{L}_{\text{IL}}(\psi) = \mathbb{E}_{\substack{(o,a)\sim\mathcal{D},\epsilon\sim\mathcal{N}(0,I) \\ k\sim\text{Uniform}\{1,\dots,K\}}} \left[||\epsilon - \epsilon_\psi(a^k_{t:t+H}, c_t, k)||^2 \right], \tag{4}$$

$$\mathcal{L}_{\text{IB}}(\theta,\phi) = \mathbb{E}_{(o,a)\sim\mathcal{D}} \left[\mathbb{E}_{\epsilon,k} \left[||\epsilon - \epsilon_\psi(a^k_{t:t+H}, \tilde{c}_t, k)||^2 \right] + \beta \text{KL}[p_\theta(Z|o)||r(Z)] \right], \tag{5}$$

where ϵ_ψ is the diffusion policy parameterized by ψ, and $a^k_{t:t+H}$ is the noisy action at diffusion step k. The overall training objective is a combination of both losses:

$$\mathcal{L}(\theta,\phi,\psi) = \mathcal{L}_{\text{IL}}(\psi) + \mathcal{L}_{\text{IB}}(\theta,\phi). \tag{6}$$

Notably, $\mathcal{L}_{\text{IB}}(\theta,\phi)$ updates only p_θ and q_ϕ, leaving the base policy unaffected, while $\mathcal{L}_{\text{IL}}(\psi)$ updates only $\mathcal{E}$ and $\mathcal{D}$. This design ensures that the exploration mechanism does not come at the cost of degrading base policy performance. Furthermore, by offloading action-level multimodal modeling to the diffusion process, the latent encoder can focus on capturing cognitive-level uncertainty, leading to a more robust characterization of the underlying task manifold.

C. User-Guided Steering

In addition to facilitating on-manifold exploration, the learned latent representation Z also provides a natural mechanism for controllable exploration. Owing to the disentanglement encouraged by $\mathcal{L}_{\text{IB}}$ (Eqn. 5), different dimensions of Z tend to correspond to distinct task-relevant factors. For example, in a cup-grasping task, one dimension may encode the cup's horizontal position, while another captures its vertical position. Such a disentangled structure allows users to intuitively guide exploration by perturbing specific latent dimensions, steering the policy toward desired behaviors.

We achieve this by first identifying the most informative dimensions of Z using a signal-to-noise ratio (SNR) criterion, and then restricting steering to these dimensions. The SNR is defined as:

$$\text{SNR}_i = \frac{\text{Var}(\mu_i)}{\mathbb{E}[\sigma_i^2]}, \quad i = 1, 2, \dots, d,$$

where μ_i and σ_i are the mean and standard deviation of the i-th dimension of Z. A high SNR indicates the dimension encodes reliable, task-relevant information, while a low SNR suggests it is uninformative or even collapsed, as validated in our experiments Sec. V-B. There is also a clear separation between effective and ineffective dimensions, allowing us to define a universal, task-agnostic threshold for categorization.

Identifying the effective dimensions substantially reduces the search space for steering, allowing users to concentrate on the most relevant factors. Furthermore, to ensure broad coverage of the task manifold, we generate a large batch of action proposals for each effective dimension and apply farthest point sampling (FPS) to select a diverse subset for user selection. By starting with a large batch size and retaining only a small number of representative actions, the selected proposals effectively capture the full range of variations along each dimension while avoiding redundancy.

These proposals are then presented to users through an interactive interface, enabling them to explore the activated latent dimensions and inspect the corresponding behaviors. By choosing preferred actions, users can directly steer the robot toward desired behaviors. This design introduces human-in-the-loop guidance during exploration, achieving even higher sample efficiency than autonomous self-improvement, while eliminating the need for specialized teleoperation devices or the exhausting effort of manual teleoperation.

V. EXPERIMENTS

In this section, we evaluate our method on a range of robot manipulation tasks in both simulation and real-world settings. We start by introducing the setups for our experiments.

A. Experiment Setups

1) Real World: We deploy our approach on a Flexiv Rizon4 robotic arm with a Robotiq 2F-85 gripper. We attach soft fingers to the grippers following [11]. The visual observations are provided by two Intel RealSense D435i cameras: one fixed on the side and one mounted on the wrist.

The robot is controlled in Cartesian space at 10 Hz. Each action chunk spans $H = 20$ steps and is represented in relative poses. The action space is 10-dimensional, comprising 3D translation, 6D rotation, and 1D gripper control, while the observation space includes RGB images ($216 \times 288 \times 3$) from both cameras, along with proprioception states.

We evaluate our method on three real-world manipulation tasks: *Mug Hang* (grasp a mug and hang it on a rack), *Toaster Load* (pick up a slice of bread and insert it into a toaster), and *Lamp Cap* (assemble an alcohol lamp by placing its cap onto its base). Fig. 3 provides an illustration of all three tasks.

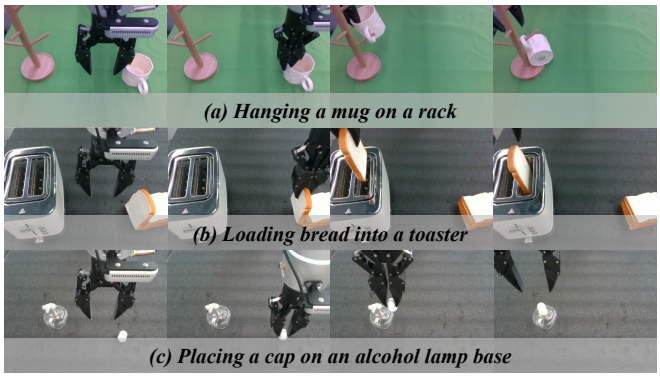


Fig. 3: **Overview of real-world manipulation tasks**, which include (a) *Mug Hang*, (b) *Toaster Load*, and (c) *Lamp Cap*.

For each task, we collect a limited set of human teleoperated demonstrations using a sigma.7 haptic device: 30 demonstrations each for *Toaster Load* and *Lamp Cap*, and 50 for *Mug Hang*, as it is relatively more complex. These demonstrations are used to train the initial base policy and serve as a foundation for subsequent self-improvement.

2) Simulation: We also evaluate our method on a set of simulation tasks from RoboMimic [34]. The tasks include *Lift*, *Can*, *Square*, and *Transport*. The observations are restricted to RGB images from a fixed agent-view camera and an in-hand camera, along with proprioceptive states.

We use the officially provided proficient-human (ph) demonstrations for policy initialization. Since Diffusion Policy already achieves near-perfect performance with the full set of 200 demonstrations [10], we consider a more challenging few-shot regime, following prior work [23]. Specifically, we randomly sample 20 demonstrations for *Can*, *Square*, and *Transport*, and 10 demonstrations for *Lift*, given its relative simplicity. Starting from these limited demonstrations and the resulting imperfect base policies, the robot is expected to improve itself through active exploration.

3) Metrics: Our primary metric is *task success rate*. To measure exploration effectiveness, we report *Pass@5*, the probability of achieving at least one success within five attempts from the same starting condition. To reflect sample efficiency, we record the *number of rollouts* required to collect these successful experiences. To assess safety and motion quality, we also compute the *average jerk* of the end-effector trajectory, defined as $J_\tau = \frac{1}{T}\sum_{t=1}^{T} ||\frac{d^3 p_t}{dt^3}||$, where p_t denotes the end-effector position at time t. A lower jerk value indicates smoother, safer motions, which is crucial for real-world deployment.

4) Policy: Our method is implemented on top of Diffusion Policy [10] with ResNet-18 [20] as the visual encoder and DDIM [42] as the scheduler. For our plug-in exploration module, both the latent encoder and decoder are 3-layer MLPs with ReLU activations, and the latent dimension is set to $d = 16$. Policies are optimized using AdamW with a learning rate of 3e-4, a batch size of 256, and trained for 25k iterations on two NVIDIA A800 GPUs.

5) Baselines: We compare our method against two baselines: the basic *Diffusion Policy (DP)* [10] without explicit exploration mechanisms, and *SIME* [23], a recently proposed approach that enables modal-level exploration by injecting random noise into diffusion conditioning. For a fair comparison, all methods are implemented to share the same base policy architecture and training recipe, differing only in their exploration mechanisms.

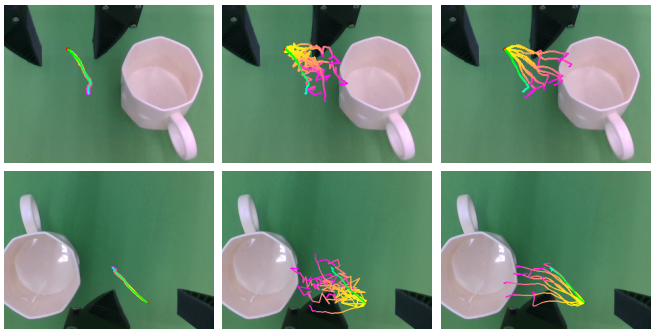
B. Evaluation on Real-World Tasks

In this section, we aim to answer the following two questions: (1) Does *SOE* lead to more effective, safer, and more efficient exploration compared to prior methods? (2) Can the learned latent representation capture task-relevant information and facilitate structured exploration as well as user-guided steering?

As shown in Table I, *SOE* consistently outperforms all baselines across three tasks in terms of effectiveness, safety, and efficiency. The basic *Diffusion Policy (DP)* exhibits low Pass@5 rates, often repeating similar unsuccessful behaviors, and can even suffer performance degradation after training on self-collected data. *SIME* achieves higher Pass@5 but produces jerky and unsafe motions, limiting the quality of the collected data and resulting in weak improvement on those precision-demanding tasks. In contrast, our approach achieves significant improvements in success rate even without human-in-the-loop steering, while maintaining smooth, safe trajectories and requiring fewer rollouts. A qualitative comparison of action proposals from different methods is shown in Fig. 4. Together, these results underscore the benefits of on-manifold exploration in enhancing the effectiveness, safety, and efficiency of robot policy self-improvement. With user-guided steering, exploration can be further directed toward promising directions, resulting in even greater improvements with fewer rollouts.

TABLE I: Experiment Results on Real-World Tasks

		Pass@5 ↑	Average Jerk ↓	Rollout Num ↓	Success Rate ↑		Relative Improvement
					Before Imp.	After Imp.	
Mug Hang	DP	0.56	3.34	77	0.47	0.38	−19.1%
	SIME	0.69	5.14 (+1.80)	65	0.47	0.50	+6.4%
	SOE (Ours)	0.75	3.57 (+0.23)	60	0.47	0.56	+19.1%
	SOE + steering (Ours)	**0.81**	3.68 (+0.34)	**53**	0.47	**0.66**	+40.4%
Toaster Load	DP	0.66	2.64	59	0.56	0.62	+10.7%
	SIME	0.84	3.64 (+1.00)	51	0.56	0.62	+10.7%
	SOE (Ours)	0.94	2.88 (+0.24)	36	0.56	0.75	+33.9%
	SOE + steering (Ours)	**1.00**	2.89 (+0.25)	**32**	0.56	**0.84**	+50.0%
Lamp Cap	DP	0.62	2.88	34	0.50	0.56	+12.0%
	SIME	0.69	3.53 (+0.65)	36	0.50	0.50	0.0%
	SOE (Ours)	0.88	2.97 (+0.09)	30	0.50	0.69	+38.0%
	SOE + steering (Ours)	**0.94**	2.90 (+0.02)	**25**	0.50	**0.81**	+62.0%



(a) DP (b) SIME (c) SOE (Ours)

Fig. 4: **Comparison of action proposals.** *DP* tends to generate repetitive failures, *SIME* explores more broadly but with erratic, temporally inconsistent motions, while our method *SOE* produces purposeful, structured proposals.

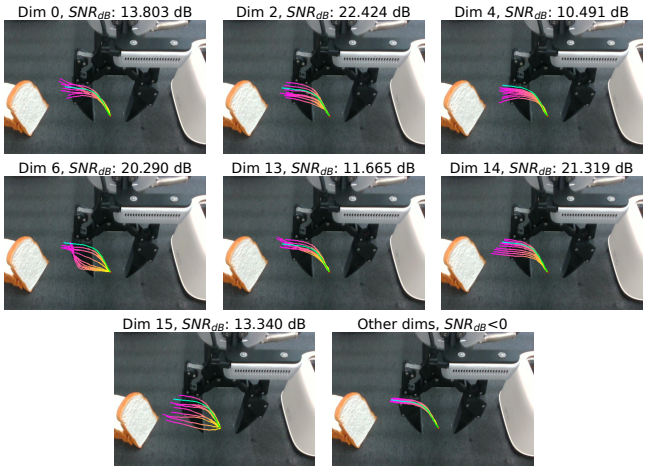


Fig. 5: **Visualization of action proposals across latent dimensions.** Each subfigure shows one dimension with its index and SNR (dB). Activating positive-SNR_{dB} dimensions yields meaningful, diverse action variations, whereas a low-SNR dimension (bottom-right) produces negligible diversity.

To gain a deeper insight into the learned latent space, we visualize the action proposals used for user-guided steering in Fig. 5. By activating one dimension at a time and plotting the resulting action proposals as trajectory points, we can observe how each dimension individually influences the robot's motion. As shown in the figure, different dimensions induce distinct patterns of action variation, such as moving left/right, up/down, or straight/curved, suggesting a certain degree of disentanglement in the latent space. Moreover, dimensions with high SNR—referred to as effective dimensions—produce meaningful trajectory variations when activated, while low-SNR dimensions yield negligible changes, indicating they contribute little to action generation. By focusing exploration and steering on the effective dimensions, our method can substantially reduce the search space, thereby resulting in more structured and efficient exploration.

C. Evaluation on Simulation Benchmark

To further investigate the robustness and scalability of *SOE*, we conduct a multi-round evaluation experiment on four vision-based manipulation tasks from RoboMimic. For each task, we train the policy under 4 different random seeds and evaluate it on 100 distinct scenarios. Each reported value is averaged over $4 \times 100 = 400$ trials, with the variation across seeds also provided to indicate the stability of policy learning. The results are shown in Fig. 6 and Fig. 7.

From the figures, we observe that our method continuously improves the policy over multiple rounds of self-improvement, achieving substantial gains in success rate with relatively few rollouts. Notably, it is the only approach that demonstrates consistent improvement across all tasks. In contrast, both baselines exhibit unstable performance, with success rates fluctuating or even declining in certain tasks. Regarding sample efficiency, although our method requires a similar number of rollouts as the baselines in the first round, it quickly reduces interaction costs in subsequent rounds, indicating that the policy effectively leverages each interaction and avoids wasting samples on uninformative exploration behaviors. Overall, these results demonstrate that our exploration mechanism can robustly improve policy performance across diverse tasks and over multiple iterations, highlighting its potential for scalable robot learning.

D. Ablation Study

To better understand the contributions of different components and the sensitivity to hyperparameters, we conduct an ablation study on the *Can* task from the RoboMimic benchmark. The results are presented in Fig. 8.

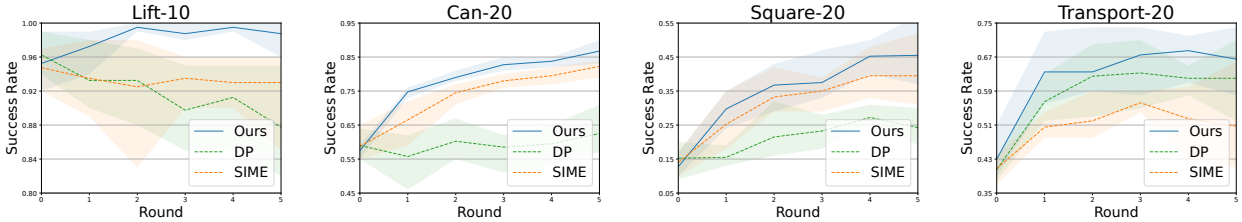


Fig. 6: **Task success rate over rounds.** Our method consistently improves policies, whereas baselines perform unstably.

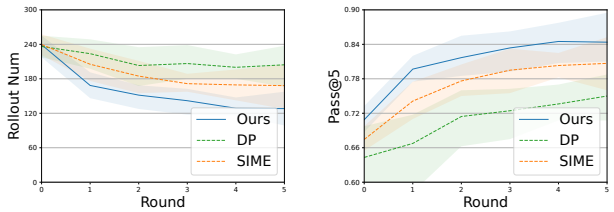


Fig. 7: **Number of rollouts and exploration success rate over rounds, averaged across tasks.** Our method generally achieves higher Pass@5 rates while requiring fewer environment interactions compared to baselines, indicating more efficient and more effective exploration.

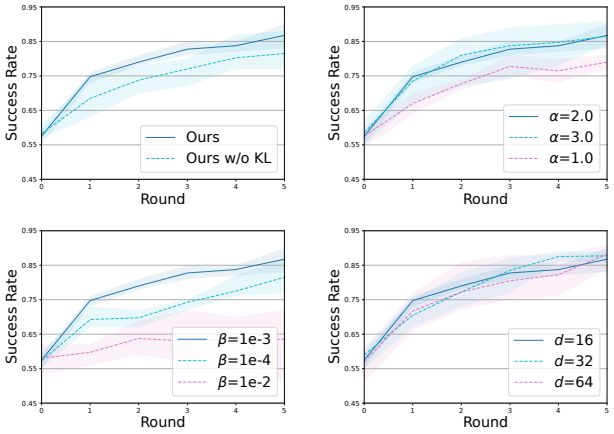


Fig. 8: **Task success rate under ablation.** We conduct these ablations on the *Can-20* task.

We begin by examining the impact of the KL term in $\mathcal{L}_{\text{IB}}$. As shown in the upper-left plot, removing the KL term reduces the final success rate from 86.75% to 81.50%, highlighting the importance of enforcing a compact and informative latent representation for effective exploration.

We then analyze the effects of the noise scale α and the KL weight β. The upper-right and bottom-left plots show that both hyperparameters play a critical role in balancing exploration diversity and stability. A small noise scale α leads to conservative exploration and sluggish improvement, whereas an excessively large α induces aggressive strategies and unstable improvement. Likewise, setting the KL weight β too low results in entangled latent representations and insufficient diversity, while too high a β can cause model collapse, failing to generate meaningful actions. To achieve a trade-off between improvement magnitude and stability, we adopt $\alpha = 2.0$ and $\beta = 0.001$ for most simulation tasks.

We further evaluate the robustness of our method to the latent dimension d. As shown in the bottom-right plot, varying d from 16 to 64 has little effect on the final success rate, suggesting that the VIB objective effectively constrains the latent space to capture only task-relevant information, independent of its nominal dimensionality. Interestingly, the number of effective dimensions, measured by SNR, remains invariant across different d. For example, in the *Can* task, the effective dimension consistently stays at 8 even when d ranges from 16 to 64. Moreover, the effective dimension appears to correlate with task complexity: simple tasks such as *Lift* require few dimensions, while more complex tasks such as *Transport* require more, as summarized in Table II.

TABLE II: **Number of Effective Dimensions.**

	Lift	*Can*	*Square*	*Transport*
Intrinsic Dim	8	8	10	16

Taken together, these results suggest the existence of a policy-agnostic intrinsic dimension for each task, which can be interpreted as the minimal degrees of freedom required to transfer task-relevant information from observations to actions. Our method is able to automatically identify and exploit these intrinsic dimensions, enabling structured on-manifold exploration that enhances both the efficiency and effectiveness of policy self-improvement.

VI. Conclusion

In this paper, we present *SOE*, a novel exploration approach for sample-efficient robot policy self-improvement. By integrating a variational information bottleneck with policy learning as an auxiliary plug-in module, our method learns a compact latent representation that enables structured on-manifold exploration, while preserving the integrity of the base policy performance. Our experiments demonstrate that this approach leads to more effective, safer, and more efficient exploration compared to prior methods. Furthermore, the learned latent space facilitates user-guided steering, allowing users to intuitively direct exploration toward desired behaviors. Overall, our work highlights the potential of structured exploration mechanisms, hoping to inspire future research in the development of sample-efficient and self-improving robot learning systems.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No.62595774), Science and Technology Major Project of Jiangsu Province (No.BG2024041), Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants.

REFERENCES

[1] Anurag Ajay et al. "Is conditional generative modeling all you need for decision-making?" In: *arXiv preprint arXiv:2211.15657* (2022).

[2] Alexander A Alemi et al. "Deep variational information bottleneck". In: *arXiv preprint arXiv:1612.00410* (2016).

[3] Lars Lien Ankile et al. "From imitation to refinement–residual rl for precise visual assembly". In: *CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data*. 2024.

[4] Johan Bjorck et al. "Gr00t n1: An open foundation model for generalist humanoid robots". In: *arXiv preprint arXiv:2503.14734* (2025).

[5] Kevin Black et al. "π_0: A Vision-Language-Action Flow Model for General Robot Control". In: *arXiv preprint arXiv:2410.24164* (2024).

[6] Konstantinos Bousmalis et al. "Robocat: A self-improving generalist agent for robotic manipulation". In: *arXiv preprint arXiv:2306.11706* (2023).

[7] Anthony Brohan et al. "Rt-1: Robotics transformer for real-world control at scale". In: *arXiv preprint arXiv:2212.06817* (2022).

[8] Yuri Burda et al. "Large-scale study of curiosity-driven learning". In: *arXiv preprint arXiv:1808.04355* (2018).

[9] Yuhui Chen et al. "Conrft: A reinforced fine-tuning method for vla models via consistency policy". In: *arXiv preprint arXiv:2502.05450* (2025).

[10] Cheng Chi et al. "Diffusion policy: Visuomotor policy learning via action diffusion". In: *The International Journal of Robotics Research* (2023), p. 02783649241273668.

[11] Cheng Chi et al. "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots". In: *arXiv preprint arXiv:2402.10329* (2024).

[12] Adrien Ecoffet et al. "Go-explore: a new approach for hard-exploration problems". In: *arXiv preprint arXiv:1901.10995* (2019).

[13] Hongjie Fang et al. "AirExo-2: Scaling up Generalizable Robotic Imitation Learning with Low-Cost Exoskeletons". In: *arXiv preprint arXiv:2503.03081* (2025).

[14] Scott Fujimoto, David Meger, and Doina Precup. "Off-policy deep reinforcement learning without exploration". In: *International conference on machine learning*. PMLR. 2019, pp. 2052–2062.

[15] Anirudh Goyal et al. "Infobot: Transfer and exploration via the information bottleneck". In: *arXiv preprint arXiv:1901.10902* (2019).

[16] Shangding Gu et al. "A review of safe reinforcement learning: Methods, theories and applications". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[17] Tuomas Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In: *International conference on machine learning*. Pmlr. 2018, pp. 1861–1870.

[18] Siddhant Haldar, Zhuoran Peng, and Lerrel Pinto. "Baku: An efficient transformer for multi-task policy learning". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 141208–141239.

[19] Siddhant Haldar et al. "Teach a robot to fish: Versatile imitation from one minute of demonstrations". In: *arXiv preprint arXiv:2303.01497* (2023).

[20] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[21] Edward S Hu et al. "Planning goals for exploration". In: *arXiv preprint arXiv:2303.13002* (2023).

[22] Michael Janner et al. "Planning with diffusion for flexible behavior synthesis". In: *arXiv preprint arXiv:2205.09991* (2022).

[23] Yang Jin et al. "SIME: Enhancing Policy Self-Improvement with Modal-level Exploration". In: *arXiv preprint arXiv:2505.01396* (2025).

[24] Hyoungseok Kim et al. "Emi: Exploration with mutual information". In: *arXiv preprint arXiv:1810.01176* (2018).

[25] Moo Jin Kim et al. "Openvla: An open-source vision-language-action model". In: *arXiv preprint arXiv:2406.09246* (2024).

[26] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. "Offline reinforcement learning with implicit q-learning". In: *arXiv preprint arXiv:2110.06169* (2021).

[27] Aviral Kumar et al. "Conservative q-learning for offline reinforcement learning". In: *Advances in neural information processing systems* 33 (2020), pp. 1179–1191.

[28] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. "Reinforcement learning with action chunking". In: *arXiv preprint arXiv:2507.07969* (2025).

[29] Timothy P Lillicrap et al. "Continuous control with deep reinforcement learning". In: *arXiv preprint arXiv:1509.02971* (2015).

[30] Huihan Liu et al. "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment". In: *The International Journal of Robotics Research* (2022), p. 02783649241273901.

[31] Jianlan Luo et al. "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning". In: *Science Robotics* 10.105 (2025), eads5033.

[32] Jianlan Luo et al. "Serl: A software suite for sample-efficient robotic reinforcement learning". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 16961–16969.

[33] Jun Lv et al. "Sam-rl: Sensing-aware model-based reinforcement learning via differentiable physics-based simulation and rendering". In: *The International Journal of Robotics Research* (2023), p. 02783649241284653.

[34] Ajay Mandlekar et al. "What matters in learning from offline human demonstrations for robot manipulation". In: *arXiv preprint arXiv:2108.03298* (2021).

[35] Suvir Mirchandani et al. "So You Think You Can Scale Up Autonomous Robot Data Collection?" In: *arXiv preprint arXiv:2411.01813* (2024).

[36] Volodymyr Mnih et al. "Playing atari with deep reinforcement learning". In: *arXiv preprint arXiv:1312.5602* (2013).

[37] Simone Parisi et al. "Interesting object, curious agent: Learning task-agnostic exploration". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20516–20530.

[38] Deepak Pathak et al. "Curiosity-driven exploration by self-supervised prediction". In: *International conference on machine learning*. PMLR. 2017, pp. 2778–2787.

[39] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 627–635.

[40] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[41] Nur Muhammad Shafiullah et al. "Behavior transformers: Cloning k modes with one stone". In: *Advances in neural information processing systems* 35 (2022), pp. 22955–22968.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[43] Richard S Sutton and Andrew G Barto. "Reinforcement Learning: An Introduction. A Bradford Book". In: *IEEE Transactions on Neural Networks* 16.1 (2005), pp. 285–286.

[44] Richard S Sutton et al. "Policy gradient methods for reinforcement learning with function approximation". In: *Advances in neural information processing systems* 12 (1999).

[45] Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

[46] Octo Model Team et al. "Octo: An open-source generalist robot policy". In: *arXiv preprint arXiv:2405.12213* (2024).

[47] Hado Van Hasselt, Arthur Guez, and David Silver. "Deep reinforcement learning with double q-learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1. 2016.

[48] Chenxi Wang et al. "Rise: 3d perception makes real-world robot imitation simple and effective". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 2870–2877.

[49] Xiu Yuan et al. "Policy decorator: Model-agnostic online refinement for large policy model". In: *arXiv preprint arXiv:2412.13630* (2024).

[50] Zheng Yuan et al. "Scaling relationship on learning mathematical reasoning with large language models". In: *arXiv preprint arXiv:2308.01825* (2023).

[51] Yanjie Ze et al. "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations". In: *arXiv preprint arXiv:2403.03954* (2024).

[52] Yu Zhang et al. "SCIZOR: A Self-Supervised Approach to Data Curation for Large-Scale Imitation Learning". In: *arXiv preprint arXiv:2505.22626* (2025).

[53] Tony Z Zhao et al. "Learning fine-grained bimanual manipulation with low-cost hardware". In: *arXiv preprint arXiv:2304.13705* (2023).

[54] Brianna Zitkovich et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control". In: *Conference on Robot Learning*. PMLR. 2023, pp. 2165–2183.