ForeDiffusion: Foresight-Conditioned Diffusion Policy via Future View Construction for Robot Manipulation

Weize Xie[1,*], **Yi Ding**[1,*], **Ying He**[1,†], **Leilei Wang**[1,2], **Binwen Bai**[1], **Zheyi Zhao**[1,2], **Chenyang Wang**[1], **F. Richard Yu**[3]

[1]College of Computer Science and Software Engineering, Shenzhen University, China
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[3]School of Information Technology, Carleton University, Canada
{2410105060, 2410103031}@mails.szu.edu.cn, heying@szu.edu.cn, wangleilei@gml.ac.cn, 2300271042@email.szu.edu.cn, zhaozheyi@gml.ac.cn, chenyangwang@ieee.org

Abstract

Diffusion strategies have advanced visual motor control by progressively denoising high-dimensional action sequences, providing a promising method for robot manipulation. However, as task complexity increases, the success rate of existing baseline models decreases considerably. Analysis indicates that current diffusion strategies are confronted with two limitations. First, these strategies only rely on short-term observations as conditions. Second, the training objective remains limited to a single denoising loss, which leads to error accumulation and causes grasping deviations. To address these limitations, this paper proposes Foresight-Conditioned Diffusion **(ForeDiffusion)**, by injecting the predicted future view representation into the diffusion process. As a result, the policy is guided to be forward-looking, enabling it to correct trajectory deviations. Following this design, ForeDiffusion employs a dual loss mechanism, combining the traditional denoising loss and the consistency loss of future observations, to achieve the unified optimization. Extensive evaluation on the Adroit suite and the MetaWorld benchmark demonstrates that ForeDiffusion achieves an average success rate of 80% for the overall task, significantly outperforming the existing mainstream diffusion methods by 23% in complex tasks, while maintaining more stable performance across the entire tasks.

Code — https://github.com/xwz-z/ForeDiffusion

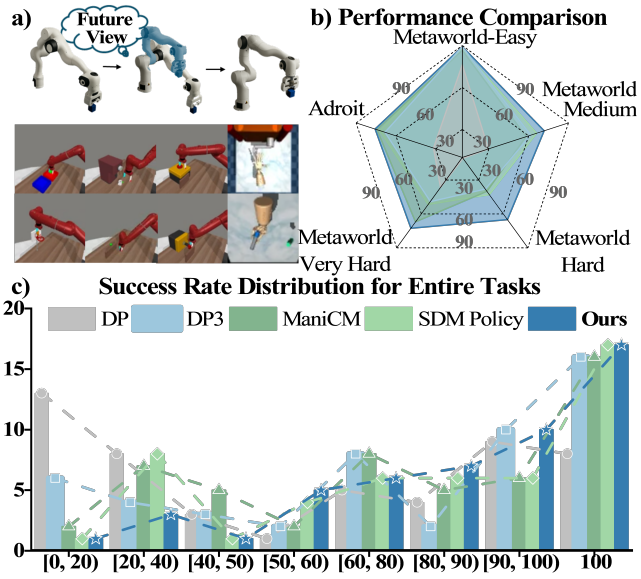

Figure 1: **Overview of ForeDiffusion.** (a) Diagram of future view-guided action generation; (b) ForeDiffusion achieves the highest average success rate across all task types; (c) ForeDiffusion shifts task counts toward the higher success rate bins across all tasks (Chi et al. 2023; Ze et al. 2024; Lu et al. 2025; Jia et al. 2024).

Introduction

Imitation learning from expert demonstrations offers a data-efficient supervised pathway to acquire diverse, task-conditioned manipulation competencies in embodied intelligence (Rahmatizadeh et al. 2018; Xie et al. 2020; Zhao et al. 2024). Building on this paradigm, diffusion-based models have become expressive visual motion strategies for robotic tasks (Urain et al. 2024). In robotic manipulation, diffusion policies have been applied in increasingly rich perceptual settings to generate action sequences with feedback control (Carvalho et al. 2023; Wu et al. 2024). Early methods conditioned on RGB inputs have achieved reliable performance from monocular observations (Chi et al. 2023). The 3D Diffusion Policy shows that compact point cloud representations significantly enhance spatial understanding across multi-task benchmarks (Ze et al. 2024). Building on these advances, diffusion-based approaches have made substantial progress in multimodal conditioning by incorporating geometric, tactile, and proprioceptive signals (Dou et al. 2024; Heng et al. 2025; Cao et al. 2024; Zhao et al. 2025).

However, current diffusion-based visuomotor policies still face significant limitations when applied to more complex manipulation tasks (Wolf et al. 2025; Xue et al. 2025). As shown in Fig. 1b, mainstream diffusion models show sig-

*These authors contributed equally.
†Corresponding author.

nificant performance degradation on complex tasks, which share the common characteristic of requiring sequential and contact-intensive interactions. (Yu et al. 2020; Lu et al. 2025). Even small control inaccuracies can accumulate over time, resulting in task failure during later stages of execution (Tsuji et al. 2025; Stepputtis et al. 2022).

We identify two primary factors causing this gap. First, current formulations typically condition diffusion on short observation sequences and assume approximate Markov sufficiency during rollout, without explicitly modeling how future scenes may evolve (Reuss et al. 2023). This limitation is especially problematic in tasks that span longer time horizons (Kim et al. 2024). Second, most policies are trained with low-level losses like denoising or single step consistency, which focus on local accuracy but provide little guidance for completing full tasks (Song et al. 2025; Fan et al. 2025). As a result, most fail with issues as misalignment, unstable grasps, or poor transitions (Ma et al. 2024).

In this paper, we address limitations stemming from statically conditioned observation horizons and single training objectives. To overcome the challenges of existing diffusion-based methods in long-horizon, complex manipulation tasks, we propose **ForeDiffusion**, a foresight-conditioned diffusion policy. ForeDiffusion tackles these challenges by constructing a compact representation of future view and injecting it into the diffusion denoising process to guide policy rollout (Fig. 1a). ForeDiffusion enables the model to reason beyond the immediate observation and mitigate cascading errors across multi-stage interactions. To support this mechanism, we introduce dual loss objective that combines standard denoising accuracy with a foresight-consistency defined over the predicted horizon. Through this design, ForeDiffusion enhances planning capabilities while ensuring stability. The proposed method surpasses recent diffusion-based baselines in overall tasks and high-success task coverage across Adroit and MetaWorld (Fig. 1b–c). Our contributions are summarized as follows:

- We introduce **ForeDiffusion**, which **injects a future view constructed** based on current observations into each denoising step, endowing diffusion policies with foresight, thus better addressing long-horizon tasks.
- We build on foresight-conditioned diffusion by formulating **dual loss objective** that combines standard denoising fidelity with prediction-consistency, thereby curbing error accumulation in complex, contact-rich manipulation.
- Compared with mainstream baselines, ForeDiffusion maintains superior overall performance and significantly **improves the success rate by 23%** in complex manipulation tasks, where other models perform poorly.

Related Work

Visuomotor Control via Diffusion Models

Diffusion models have emerged as a powerful paradigm for generating action sequences in robotic manipulation, leveraging iterative denoising processes inspired by their success in image and video synthesis (Ho, Jain, and Abbeel 2020; Ho et al. 2022). Early works like Diffusion Policy are conditioned on RGB observations to achieve reliable performance in single-arm manipulation tasks, outperforming traditional behavior cloning methods (Chi et al. 2023). Recent advances incorporate richer perceptual inputs, such as 3D Diffusion Policy (DP3), which uses point cloud representations to enhance spatial reasoning and data efficiency on multi-task benchmarks like MetaWorld (Ze et al. 2024). Other innovations include FlowPolicy, which employs manifold-aligned denoising to capture low-dimensional action structures (Zhang et al. 2025a), ManiCM integrates multimodal inputs into a consistency-driven 3D diffusion framework for real-time inference (Lu et al. 2025), and SDM Policy, which uses teacher-student distillation to enable single-step generation, improving task success rates (Jia et al. 2024). Recent surveys further categorize diffusion policies in grasp learning, trajectory planning, and skill acquisition (Zhang et al. 2025b; Ma et al. 2024). However, these approaches typically rely on static conditioning based primarily on initial proprioceptive inputs, without modeling temporal evolution, which limits foresight and leads to degraded performance in dynamic or complex tasks (Lv et al. 2025; Reuss et al. 2023).

Foresight-Driven Control with Dual Loss Training

A parallel line of research studies predictive world models that forecast future observations, giving the agent foresight beyond its current sensory window (Hafner et al. 2019). Latent video predictor DreamerV3 rolls out imagined trajectories in a learned latent space to evaluate long-horizon returns (Hafner et al. 2023). MoDem-V2 integrates RGB-based perception with learned visuo-motor models to perform contact-rich manipulation in uninstrumented, real-world environments (Lancaster et al. 2024). Hierarchical Diffusion Policy (HDP) and Causal Diffusion Policy (CDP) split prediction across coarse and fine temporal scales, allowing long range forecasts to steer fine-grained actions (Ma et al. 2024, 2025). Evidence from other fields confirms the benefit of explicit foresight. SceneDiffuser and MotionDiffuser add cost or safety terms to diffusion based trajectory forecasts, lowering collision rates on Waymo and nuScenes (Jiang et al. 2023, 2024). DiffuseLoco stabilizes quadruped gaits by penalizing high energy latent rollouts, while MVDiffusion and Percept-Diff pair pixel denoising with perceptual to keep long videos semantically coherent (Huang et al. 2025; Shi et al. 2023; Borno et al. 2024).

To stabilize training and reduce exposure bias, many works add task level consistency losses to the basic reconstruction or denoising objective (Li et al. 2024). Representative examples include the self predictive loss in TD-MPC2 (Hansen, Su, and Wang 2023), horizon wise KL regularization in Trajectory Transformer++ (Janner, Li, and Levine 2021), and the score plus value distillation in SDMP (Jia et al. 2024). Collectively, prior work shows that combining explicit future prediction with joint local and global supervision leads to more reliable control on complex, long-horizon tasks (Chen et al. 2024).

Building on 3D Diffusion Policy, to tackle the performance decline of complex tasks caused by limited perception and weak task supervision, we propose **ForeDiffusion**, a modular, foresight-aware visuomotor policy. By conditioning each diffusion step on a compact, predicted future

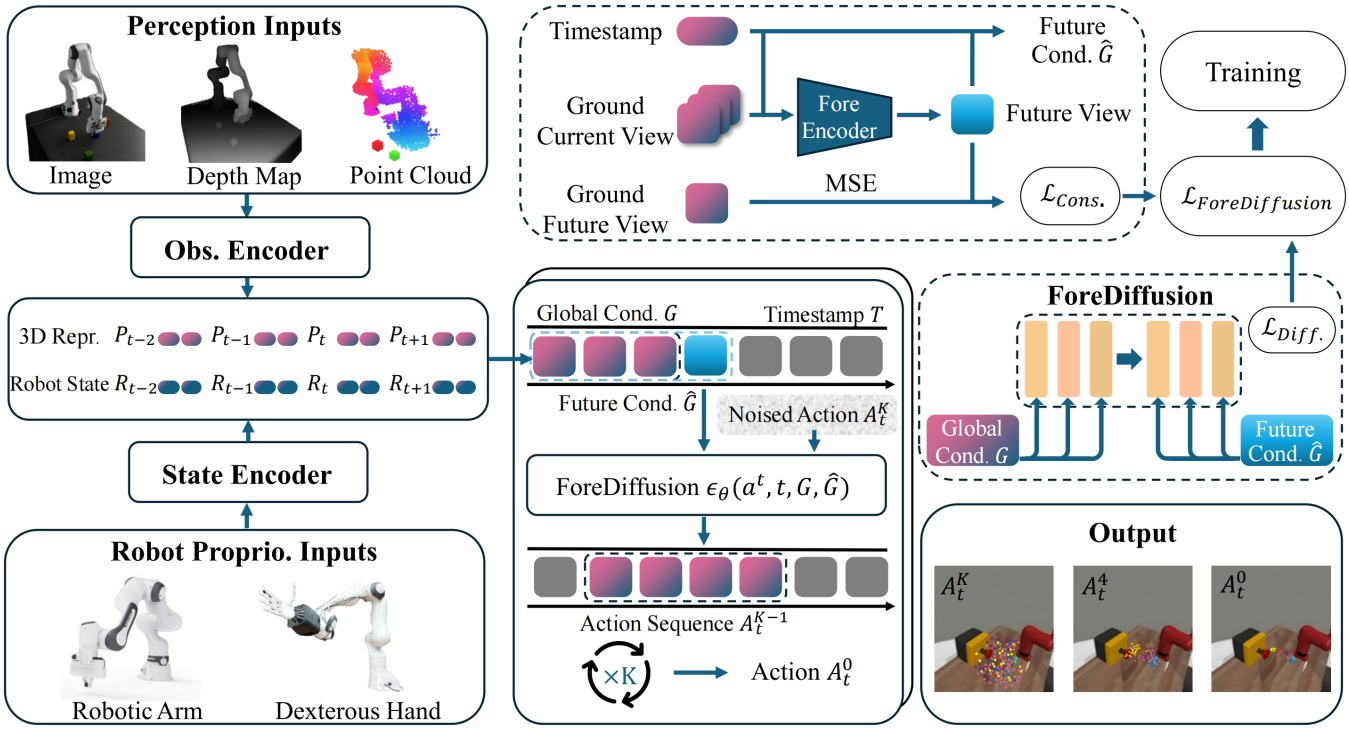

Figure 2: **Architecture of ForeDiffusion.** The perception module fuses RGB-D and proprioceptive inputs into 3D latent representations; an observation encoder outputs a global condition G and a future condition $\hat{G}$. These condition guide a K-step reverse-diffusion process that denoises a noise-perturbed action trajectory into an executable sequence A_0, with joint construction and behavioral losses enforcing accurate future prediction and expert-level control.

view and optimizing with a dual loss objective that balances construction fidelity and long-horizon consistency, ForeDiffusion delivers more stable and anticipatory control. Experiments show that while it matches mainstream methods on basic tasks, it consistently surpasses them on complex, contact-rich manipulation tasks, achieving markedly higher success rates under increasing task complexity.

Method

In this section, we first describe how to construct a future view from observations to approximate the future perception. Then, we develop how the current and future views are combined as conditions to guide the denoising process and generate trajectories that align with expert intent. Finally, we introduce a dual loss mechanism that jointly supervises both the future view construction and the denoising objectives.

Future View Construction

To enable the model to reason about future outcomes while relying on present information at inference, we introduce a future view construction mechanism that predicts a forward-looking representation from near-term observations. At each time step t, we define the current observation as a pair of temporally adjacent frames, $O_t^{cur.} = (O_{t-1}, O_t)$, where each $O_t = (P_t, R_t)$ consists of a 3D point cloud P_t and proprioceptive robot state R_t (Janner et al. 2022). We encode this pair using a shared observation encoder $Enc(\cdot)$ to obtain the current observation $F_t^{cur.}$, defined as:

$$F_t^{cur.} = Enc\left((P_{t-1}, R_{t-1}), (P_t, R_t)\right) \tag{1}$$

The ground-truth future view $F_t^{gt.}$ is constructed by the observation at time $t+1$. Based on the current observation feature $F_t^{cur.}$, we construct a predictive future view $F_t^{cons.}$ using a multilayer perceptron (MLP). The MLP learns to map the current observation to a future scene representation, enabling the model to predict future view based on near-term observations.

Foresight-Conditioned Denoising

The global condition $\mathbf{G}$ and the future condition $\hat{\mathbf{G}}$ denote separately the temporally-aware representations constructed by concatenating $F_t^{cur.}$ and $F_t^{cons.}$ with a timestamp encoding and enables the model to incorporate timing information into both the current and predicted future view. Our diffusion model generates actions by reversing a learned conditional diffusion process (Ho, Jain, and Abbeel 2020). Starting from an initial Gaussian noise $\mathbf{a}^T \sim \mathcal{N}(0, I)$, the model iteratively refines the noisy action over T denoising steps (see Fig. 3). Each step is performed as follows:

$$\begin{aligned}\mathbf{a}^{t-1} &= \text{Denoise}(\alpha_t, \sigma_t, \mathbf{a}^t, t, \mathbf{G}), \\ &= \alpha_t \left(\mathbf{a}^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \epsilon_\theta(\mathbf{a}^t, t, \mathbf{G}, \hat{\mathbf{G}})\right) \\ &\quad + \sigma_t \cdot \mathbf{z}, \quad z \sim \mathcal{N}(0, I), \quad t = T, \ldots, 1. \end{aligned} \tag{2}$$

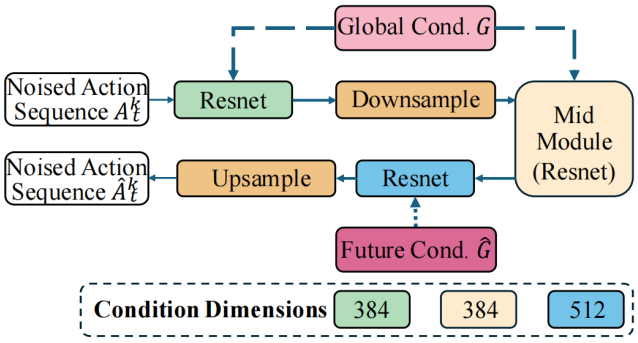

Figure 3: **Architecture of Foresight-Diffusion.** A ResNet encoder–decoder injects 384 / 512-dimensional context vectors G and $\hat{G}$ across all sampling stages to denoise the action token A_k^t into $\hat{A}_k^t$, with anticipated future observations.

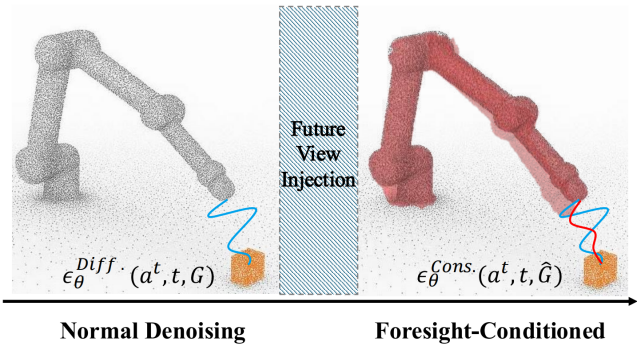

Figure 4: **Schematic of Future View Injection.** Normal denoising performs the standard early-phase diffusion, whereas foresight conditioning injects future-view information, allowing the model to anticipate consecutive outcomes and generate more stable, goal-aligned action sequences.

where ϵ_θ denotes a conditional denoising network that predicts the noise added to the action $\mathbf{a}^t$, given the timestep t, global condition $\mathbf{G}$, and future condition $\hat{\mathbf{G}}$. The denoising process is modulated by a variance schedule $\{\alpha_t, \bar{\alpha}_t, \sigma_t\}$, and the final output $\mathbf{a}^0$ serves as the predicted action.

This iterative procedure can be interpreted as a gradient-based update within an implicit energy field over the action space. Specifically, the denoising step resembles a descent in a learned energy field $E(x)$, yielding a simplified rule:

$$\mathbf{a}' = \mathbf{a} - \gamma_t \cdot \nabla E(x) \tag{3}$$

where γ_t controls the step size and $\nabla E(x)$ represents the direction for refining the noisy action. While this abstraction omits scheduling details, it offers an intuitive view on how the diffusion model guide actions toward more optimal regions in the behavior space, shaped by current observations and predicted future view.

Dual Diffusion Loss

To ensure that this constructed feature captures forward-looking information, we supervise it using a mean squared error loss against the encoded ground-truth future view:

$$\mathcal{L}_{Construction} = \left\| F_t^{cons.} - F_t^{gt.} \right\|_2^2 \tag{4}$$

In parallel, the diffusion model is trained to reverse the noise process with a denoising objective:

$$\mathcal{L}_{Diffusion} = E_{x_t,t}\left[\left\| \epsilon_\theta(\mathbf{a}^t, t, \cdot) - \epsilon \right\|_2^2 \right] \tag{5}$$

The total training objective is a weighted sum of the diffusion and structure alignment losses:

$$\mathcal{L}_{ForeDiffusion} = \mathcal{L}_{Diff.} + \beta \cdot \mathcal{L}_{Cons.} \tag{6}$$

where β controls the influence of the prediction alignment.

$$\mathbf{a}^{t-1} = \mathbf{a}^t - \gamma_t \cdot \left(\epsilon_\theta^{Diff.}(\cdot) + \beta \cdot \epsilon_\theta^{Cons.}(\cdot) \right) \tag{7}$$

This design enables the model to construct and utilize a structured predictive future view based solely on near-term observations, providing a form of implicit foresight that generalizes to inference without requiring access to future data.

In summary, at each time step t, ForeDiffusion extracts the current observation and predicts the future representation to form the diffusion conditions $(\mathbf{G}, \hat{\mathbf{G}})$ that guide each denoising step (see Fig. 4). Given expert trajectories, the goal is to learn a policy $\epsilon(a^t|O_{1:t})$ that mimics expert behavior. The training is supervised by a dual loss objective, combining the standard diffusion loss $\mathcal{L}_{Diff.}$ with an structure loss $\mathcal{L}_{Cons.}$ to improve both stability and fidelity.

Intropy Analysis of ForeDiffusion

The **Intropy framework** can be used to quantify intelligence as $d\mathcal{L} = \frac{\delta S}{R}$, where $d\mathcal{L}$ (Intropy) denotes the incremental intelligence gain, δS represents newly absorbed information, and R reflects the system's internal state, such as complexity or uncertainty (Ren et al. 2025).

Under this view, **ForeDiffusion** maximizes *intropy efficiency* by increasing information gain (δS) while reducing structural resistance (R). Its foresight-conditioned denoising, guided by predicted future views F_t^{cons}, injects anticipatory information that amplifies long-horizon learning signals. The dual loss $\mathcal{L} = \mathcal{L}_{\text{Diff}} + \beta \cdot \mathcal{L}_{\text{Cons}}$, further stabilizes representation and reduces uncertainty. Through the Intropy lens, ForeDiffusion achieves greater $d\mathcal{L}$—transforming predictive entropy into structured foresight—yielding robust, intelligent visuomotor control with an average success rate of 80.6% on long-horizon, contact-rich manipulation tasks.

Experiments

In this section, we conduct the experiments that focus on six research questions: **(RQ1)** Does the ForeDiffusion outperform baselines in task success rate? **(RQ2)** Is it effective for improving the learning efficiency across different tasks? **(RQ3)** As task complexity increases, does the performance remain consistently robust? **(RQ4)** Is foresight injection position essential for performance enhancement? **(RQ5)** How sensitive is the weighting of the dual loss during training?

Methods	Adroit Benchmark			MetaWorld Benchmark				Average
	Hammer	Pen	Door	Easy	Medium	Hard	Very Hard	
DP (Chi et al. 2023)	45±5	13±2	37±2	82±26	31±24	11±13	37±27	56.64±37.61
FlowPolicy (Zhang et al. 2025a)	100±0	53±12	58±5	80±34	62±32	49±40	36±7	62.57±21.00
ManiCM (Lu et al. 2025)	100±0	48±3	**75±3**	86±23	58±25	40±32	67±26	71.83±30.37
SDM Policy (Jia et al. 2024)	100±0	49±4	68±1	89±19	66±28	43±37	52±32	74.81±30.18
DP3 (Ze et al. 2024)	100±0	43±6	62±4	**91±19**	61±32	44±41	49±33	72.35±32.77
ForeDiffusion (Ours)	**100±0**	**59±12**	68±1	89±16	**73±26**	**59±29**	**75±28**	**80.56±22.93** (↑ 5.75%)

Table 1: **Average success rates (%) of ForeDiffusion and baselines on the Adroit and MetaWorld benchmarks.** ForeDiffusion achieves the top average success rate (80.56%), matching or surpassing all baselines on every Adroit task and dominating the Medium, Hard, and Very Hard tiers of MetaWorld by margins of 6–26%. Its consistently superior scores across the most challenging settings underscore the benefit of foresight-conditioned diffusion for robust grasping and long-horizon planning.

(RQ6) How does the scale of demonstrations influence performance and scalability?

Experiment Setups

Simulation Benchmark To evaluate our method, we adopt two representative simulation platforms as benchmarks: Adroit (Rajeswaran et al. 2017) and MetaWorld (Yu et al. 2020). Adroit offers high-DoF dexterous hand manipulation with complex motor control; MetaWorld is implemented using the MuJoCo physics simulator (Todorov, Erez, and Tassa 2012). It provides more robot operation tasks and serves as a standard benchmark for evaluating policy generalization and task transferability. Furthermore, as the length and complexity of the task time series increase, they are often categorized as Easy/Medium/Hard/Very Hard (McLean et al. 2025; Seo et al. 2023). These benchmarks cover the entire process from low-level control to high-level perceptual reasoning. To unify evaluation metrics, based on the MetaWorld classification criteria, we further termed the Medium, Hard, and Very Hard tiers of MetaWorld as complex tasks (Hu, Mirchandani, and Sadigh 2023).

Expert Demonstrations For expert demonstrations, we adopt domain-specific strategies to ensure high-quality and consistent supervision across tasks. In MetaWorld, expert trajectories are generated using built-in scripted policies. For other domains, expert data is collected from reinforcement learning agents trained to solve the tasks (Nguyen and La 2019; Schulman et al. 2017). Specifically, for Adroit, we employ VRL3 (Wang et al. 2022) to obtain successful trajectories. We ensure that all imitation learning algorithms are trained with the same set of expert trajectories.

Baselines We compare against five representative diffusion based baselines. Diffusion Policy (Chi et al. 2023) formulates single-stage conditional action generation. DP3 (Ze et al. 2024) extends this with 3D-conditioned modeling for visuomotor control. FlowPolicy (Zhang et al. 2025a) enforces trajectory coherence via implicit temporal modeling. ManiCM (Lu et al. 2025) integrates multimodal inputs into a consistency-driven 3D diffusion framework. SDM Policy (Jia et al. 2024) adopts teacher–student distillation for accelerated inference. These baselines span a spectrum of design philosophies from flat to structured diffusion, and from explicit to implicit behavior modeling.

Implementation Details We implement our method built upon a conditional U-Net backbone. FiLM-style (Perez et al. 2018) conditional modulation is applied throughout the network, with asymmetric conditioning dimensions: 384 for the downsampling and 512 for the upsampling. Point clouds are encoded via a PointNet-style encoder (Qi et al. 2017) with 3 input channels and 64 output dimensions. We train the policy using AdamW with a learning rate of 1e-4, cosine learning rate schedule, and 500 warm-up steps. The model is trained for 3000 epochs with a batch size of 128. Diffusion is implemented with the DDIM scheduler (Song, Meng, and Ermon 2020), and EMA is applied to stabilize training. All experiments are performed on a single NVIDIA RTX 3080 GPU.

Evaluation Metrics Based on the evaluation protocol of DP3, each experiment is conducted with 3 random seeds (0, 1, 2). Throughout training, the policy is evaluated every 200 epochs using 15 rollouts over a total of 3,000 episodes. For each seed, we compute the average of the top 5 success rates across all evaluations, and report the final performance as the mean and standard deviation over the 3 seeds.

Comparison with State-of-the-art Methods

Success Rate (RQ1) We compare the success rates of our method with state-of-the-art baselines on Adroit suits and MetaWorld benchmark. Our method consistently achieves competitive performance, particularly excelling in complex, dexterous and multi-task settings. Table 1 compares policy performance across the Adroit and MetaWorld benchmarks. On the Adroit tasks, ours achieves 100% on Hammer, and also surpasses prior methods on Pen and Door. In MetaWorld, ForeDiffusion maintains strong results as task complexity increases. On Medium, Hard and Very Hard subsets, it reaches 73%, 59% and 75% success rates, respectively, with standard deviations of 26, 29 and 28. These results represent gains of up to 26% points over the best-performing baseline, and are particularly significant on Hard and Very Hard tasks where most baselines degrade sharply. FlowPolicy drops to 36% on Very Hard, while DP3 achieves only 44% on Hard. Overall, ForeDiffusion delivers an average success rate of 80.6% with a standard deviation of 22.9 across all tasks, outperforming the strongest baseline

Methods	Medium				Hard			Very Hard					Average
	B	BP	H	PW	HI	POH	P	D	PPW	SPe	SPl	SPh	
Diffusion Policy (DP)	85	15	15	20	9	0	30	43	5	11	11	63	26
3D Diffusion Policy (DP3)	98	34	76	49	14	14	51	69	35	17	27	97	48
ForeDiffusion (Ours)	**100**	**39**	**97**	**99**	**20**	**53**	**75**	**94**	**92**	**39**	**49**	**100**	**71** (↑ 23%)

Table 2: **Success rates (%) of different policies on challenging metaworld tasks.** ForeDiffusion delivers the highest score on every one of the complex tasks in MetaWorld, raising the overall mean success to 71%, a gain of 23% over DP3 and 45% over the Diffusion Policy. The advantages are most evident on the Very Hard tasks, up to 92% on PPW and 100% on SPh. These results confirm that ForeDiffusion markedly improves robustness under increasing task complexity.

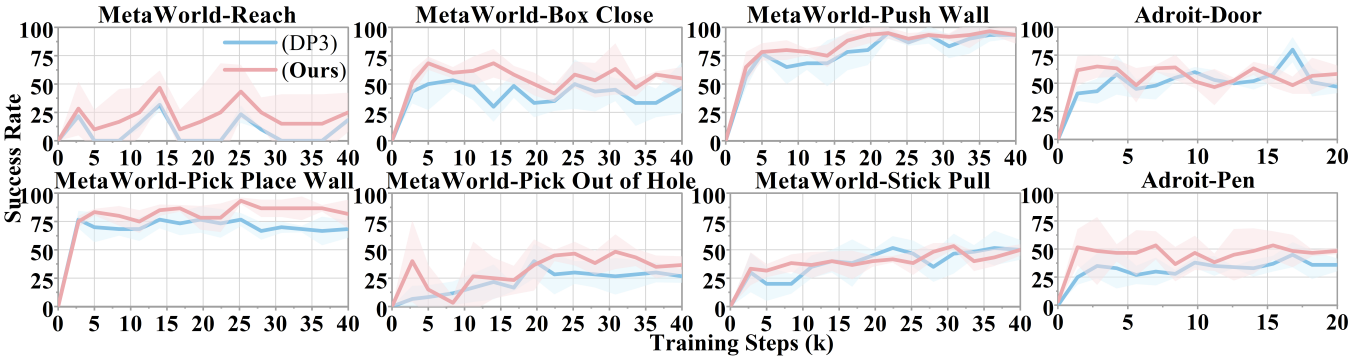

Figure 5: **Learning efficiency.** Compared to DP3, ForeDiffusion shows higher stability, learning efficiency, and success rates.

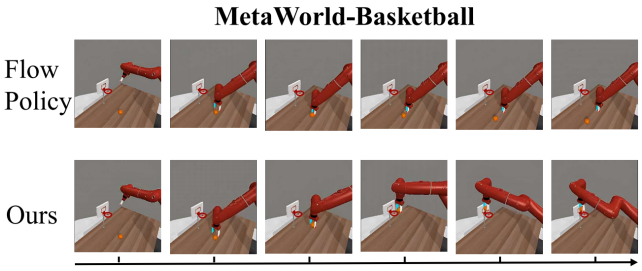

Figure 6: **Basketball task on different method.** ForeDiffusion consistently scores, whereas FlowPolicy misses, demonstrating superior grasp stability, trajectory planning.

SDM Policy (74.8%). Overall, the results underline ForeDiffusion's superior success rates and its robust, steady performance on the most demanding manipulation tasks.

Learning Efficiency (RQ2) To evaluate learning efficiency, we compare the success rate curves of different methods across tasks of varying complexity (see Fig. 5). Our method consistently outperforms DP3 under low-data regimes. On Disassemble, we achieve 95% success with only 10 demonstrations, while DP3 reaches just 40%. For harder tasks like Shelf Place, we reach 55% with 10 demonstrations, whereas DP3 remains at 25%. Even in tasks where DP3 eventually catches up (e.g. Stick Push), our method converges faster (100% at 10 demos vs. 90% for DP3 at 20). Overall, our policy achieves over 90% success on 4 out of 5 tasks within 20 demonstrations, while DP3 requires 50 demonstrations to reach comparable performance.

Prospective Learning Ability (RQ3) To evaluate the model adaptability and reasoning across rising task complexity, we evaluate Foresight-Conditioned Diffusion on 12 MetaWorld tasks spanning the Medium, Hard, and Very Hard tasks in the MetaWorld benchmark: Basketball (**B**), Bin Picking (**BP**), Hammer (**H**), Push Wall (**PW**), Hand Insert (**HI**), Pick Out of Hole (**POH**), Push (**P**), Disasemble (**D**), Pick Place Wall (**PPW**), Shelf Place (**SPe**), Stick Pull (**SPl**), Stick Push (**SPh**). As shown in Table 2, ours consistently surpasses all baselines, with the margin widening at complex tasks. This demonstrates the foresight module allows the policy to exploit future information, anticipate long-horizon outcomes, and make more robust decisions capabilities that standard diffusion policies lose under high uncertainty or long temporal dependencies. Fig. 6 compares FlowPolicy with our method on the same simulated task, confirming that ForeDiffusion excels on challenges demanding long-term reasoning and planning.

Ablations

Foresight-Conditioned Diffusion (RQ4) To evaluate the effectiveness of our method, we conduct ablation studies focusing on the point at which future view is injected into the diffusion process. In the baseline, future view is entirely removed, and the model relies solely on current view, allowing us to assess the contribution of foresight-conditioned diffusion itself. In the early-stage, future view are fused at the input stage of the U-Net alongside global conditioning, influencing the entire diffusion process from the beginning. In contrast, ForeDiffusion introduces future view at the mid-stage of the U-Net, directly altering the downstream gener-

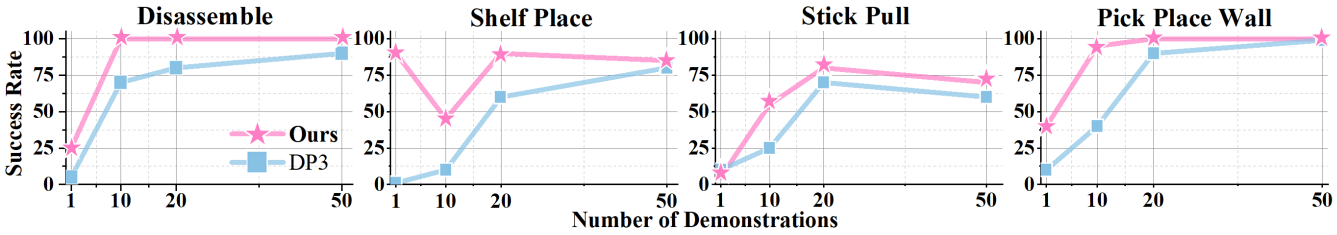

Figure 7: **Qualitative comparison of efficient scaling with demonstrations.** To evaluate the impact of demonstration quantity on policy performance, we conduct experiments on four Very Hard tasks from the MetaWorld. We compare ForeDiffusion and DP3 by progressively increasing the number of expert demonstrations during training. Results show that ForeDiffusion consistently achieves higher success rates than DP3 under low-data regimes, demonstrating strong sample efficiency.

Injection Position	MetaWorld Benchmark Medium	Hard	Very Hard	**Average**
w/o Future view	61±32	44±41	49±33	50.5±32.6
Early (Ours)	71±29	**60±29**	66±31	67.7±28.1
Mid-Stage (Ours)	**73±26**	59±29	**75±28**	**70.3±27.3**

Table 3: **Ablation results on the MetaWorld compare different future view injection position.** Introducing foresight from the mid-stage of the diffusion leads to the best overall performance across Medium and Very Hard tasks.

Dual Loss Weight	MetaWorld Benchmark Medium	Hard	Very Hard	**Average**
w/o Dual Loss	71±29	58±32	66±30	67.8±28.6
Dynamic (Ours)	70±30	**62±31**	69±30	68.3±28.7
Fixed (Ours)	**73±26**	59±29	**75±28**	**70.3±27.3**

Table 4: **Ablation studies on the effect of dual-loss strategies under different difficulty levels in MetaWorld.** The fixed weight dual loss achieves the best performance on Medium and Very Hard tasks and the highest average score.

ative trajectory while preserving the early-stage representation learning (see Table 3). Experimental results show that removing foresight leads to significantly worse performance and reduced policy stability, while early injection offers limited gains due to the dilution of future context. Our mid-stage injection achieves the best performance, validating that conditioning diffusion at structurally critical stages enables more effective use of future view.

Dual Loss Synergy (RQ5) In order to verify the effectiveness of the proposed dual loss mechanism, we design three groups of comparative experiments: the first group completely removes the auxiliary loss and only use the diffusion prediction loss for training; the second group adopts a dynamic weight strategy, in which the proportion of the auxiliary loss increases with training steps, following an exponential growth, allowing the model to progressively adjust its reliance on supervision signals at different stages; the third group is the main method of this paper, which uses fixed weights to fuse the main loss and auxiliary loss. To evaluate the impact of the dual loss design, we perform ablation studies on the MetaWorld benchmark under varying task complexities. As shown in Table 4, removing the dual loss significantly degrades performance, especially on Very Hard tasks, suggesting that dual loss plays a critical role in guiding the diffusion process. Although dynamically weighting the two losses improves results, it still suffers from instability across tasks. In contrast, our fixed weighting dual loss formulation achieves the best overall performance, reaching an average success rate of 70.3%. This highlights the effectiveness and stability of incorporating future view supervision in a balanced and consistent manner.

Scaling with Expert Demonstrations (RQ6) We evaluate the scalability of different methods with respect to the number of expert demonstrations, focusing on four Very Hard tasks from MetaWorld. As shown in Fig. 7, our method consistently outperforms DP3 across all demonstration settings. Remarkably, even with as few as a single demonstration, our method achieves strong performance, reaching success rates above 75% on some tasks. This indicates that our method is able to extract and generalize task-relevant information efficiently, owing to the inductive structure imposed by foresight representation learning. As the number of demonstrations increases, our performance continues to improve steadily, widening the gap with baseline methods.

Conclusion

In this work, we propose ForeDiffusion, a foresight conditioned diffusion policy that predicts future view and optimizes long-horizon consistency, overcoming the short-horizon conditioning and single loss constraints of existing approaches, thereby preventing the sharp drop in success rate that occurs as task complexity rises. A compact future-view latent is injected at the diffusion process, and a dual-loss design balances denoising fidelity with trajectory-level coherence to keep performance stable over extended horizons. On the Adroit and MetaWorld benchmarks, it achieves an average success rate of 80.6% and surpasses 3D Diffusion Policy by up to 23% in complex manipulation tasks, demonstrating strong generalization, data efficiency, and interpretability. We hope that ForeDiffusion drives further exploration to improve the performance of diffusion strategies towards manipulation applications.

Acknowledgments

This work is supported in part by Shenzhen Science and Technology Program under Grant ZDSYS20220527171400002, the National Natural Science Foundation of China (NSFC) under Grants 62271324, 62231020 and 62371309.

References

Borno, M. N. A.; Raihan, M. T.; Ahmed, A.; Shovon, M. S. H.; Shin, J.; and Mridha, M. 2024. Percept-diff: Innovations in stable diffusion for high-fidelity ihc image generation in her2 breast cancer incorporating perceptual loss. In *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 520–526. IEEE.

Cao, J.; Liu, J.; Kitani, K.; and Zhou, Y. 2024. Multi-Modal Diffusion for Hand-Object Grasp Generation.

Carvalho, J.; Le, A. T.; Baierl, M.; Koert, D.; and Peters, J. 2023. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1916–1923. IEEE.

Chen, C.; Deng, F.; Kawaguchi, K.; Gulcehre, C.; and Ahn, S. 2024. Simple Hierarchical Planning with Diffusion. In *The Twelfth International Conference on Learning Representations*. The International Conference on Learning Representations (ICLR).

Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; and Song, S. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*.

Dou, Y.; Yang, F.; Liu, Y.; Loquercio, A.; and Owens, A. 2024. Tactile-augmented radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26529–26539.

Fan, S.; Yang, Q.; Liu, Y.; Wu, K.; Che, Z.; Liu, Q.; and Wan, M. 2025. Diffusion trajectory-guided policy for long-horizon robot manipulation.

Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J. 2019. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, 2555–2565. PMLR.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2023. Mastering diverse domains through world models.

Hansen, N.; Su, H.; and Wang, X. 2023. Td-mpc2: Scalable, robust world models for continuous control.

Heng, L.; Geng, H.; Zhang, K.; Abbeel, P.; and Malik, J. 2025. ViTacFormer: Learning Cross-Modal Representation for Visuo-Tactile Dexterous Manipulation.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in neural information processing systems*, 35: 8633–8646.

Hu, H.; Mirchandani, S.; and Sadigh, D. 2023. Imitation bootstrapped reinforcement learning. *arXiv preprint arXiv:2311.02198*.

Huang, X.; Chi, Y.; Wang, R.; Li, Z.; Peng, X. B.; Shao, S.; Nikolic, B.; and Sreenath, K. 2025. DiffuseLoco: Real-Time Legged Locomotion Control with Diffusion from Offline Datasets. In *Conference on Robot Learning*, 1567–1589. PMLR.

Janner, M.; Du, Y.; Tenenbaum, J.; and Levine, S. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning*, 9902–9915. PMLR.

Janner, M.; Li, Q.; and Levine, S. 2021. Offline Reinforcement Learning as One Big Sequence Modeling Problem. In *Advances in Neural Information Processing Systems*.

Jia, B.; Ding, P.; Cui, C.; Sun, M.; Qian, P.; Huang, S.; Fan, Z.; and Wang, D. 2024. Score and Distribution Matching Policy: Advanced Accelerated Visuomotor Policies via Matched Distillation.

Jiang, C.; Cornman, A.; Park, C.; Sapp, B.; Zhou, Y.; Anguelov, D.; et al. 2023. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9644–9653.

Jiang, M.; Bai, Y.; Cornman, A.; Davis, C.; Huang, X.; Jeon, H.; Kulshrestha, S.; Lambert, J.; Li, S.; Zhou, X.; et al. 2024. Scenediffuser: Efficient and controllable driving simulation initialization and rollout. *Advances in Neural Information Processing Systems*, 37: 55729–55760.

Kim, S.; Choi, Y.; Matsunaga, D. E.; and Kim, K.-E. 2024. Stitching sub-trajectories with conditional diffusion model for goal-conditioned offline rl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13160–13167.

Lancaster, P.; Hansen, N.; Rajeswaran, A.; and Kumar, V. 2024. Modem-v2: Visuo-motor world models for real-world robot manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 7530–7537. IEEE.

Li, H.; Zhang, Y.; Wen, H.; Zhu, Y.; and Zhao, D. 2024. Stabilizing diffusion model for robotic control with dynamic programming and transition feasibility. *IEEE Transactions on Artificial Intelligence*, 5(9): 4585–4594.

Lu, G.; Gao, Z.; Chen, T.; Dai, W.; Wang, Z.; Ding, W.; and Tang, Y. 2025. ManiCM: Real-time 3D Diffusion Policy via Consistency Model for Robotic Manipulation. arXiv:2406.01586.

Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M. Y.; and Nie, L. 2025. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17394–17404.

Ma, J.; Qin, Y.; Li, Y.; Liao, X.; Guo, Y.; and Zhang, R. 2025. CDP: Towards Robust Autoregressive Visuomotor Policy Learning via Causal Diffusion.

Ma, X.; Patidar, S.; Haughton, I.; and James, S. 2024. Hierarchical diffusion policy for kinematics-aware multi-task

robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18081–18090.

McLean, R.; Chatzaroulas, E.; McCutcheon, L.; Röder, F.; Yu, T.; He, Z.; Zentner, K.; Julian, R.; Terry, J.; Woungang, I.; et al. 2025. Meta-World+: An Improved, Standardized, RL Benchmark. *arXiv preprint arXiv:2505.11289*.

Nguyen, H.; and La, H. 2019. Review of deep reinforcement learning for robot manipulation. In *2019 Third IEEE international conference on robotic computing (IRC)*, 590–595. IEEE.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Rahmatizadeh, R.; Abolghasemi, P.; Bölöni, L.; and Levine, S. 2018. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. In *2018 IEEE international conference on robotics and automation (ICRA)*, 3758–3765. IEEE.

Rajeswaran, A.; Kumar, V.; Gupta, A.; Vezzani, G.; Schulman, J.; Todorov, E.; and Levine, S. 2017. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations.

Ren, Y.; Zhang, H.; Yu, F. R.; Li, W.; Zhao, P.; and He, Y. 2025. Industrial Internet of Things With Large Language Models (LLMs): An Intelligence-Based Reinforcement Learning Approach. *IEEE Transactions on Mobile Computing*, 24(5): 4136–4152.

Reuss, M.; Li, M.; Jia, X.; and Lioutikov, R. 2023. Goal-conditioned imitation learning using score-based diffusion policies.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms.

Seo, Y.; Hafner, D.; Liu, H.; Liu, F.; James, S.; Lee, K.; and Abbeel, P. 2023. Masked world models for visual control. In *Conference on Robot Learning*, 1332–1344. PMLR.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models.

Song, M.; Deng, X.; Zhou, Z.; Wei, J.; Guan, W.; and Nie, L. 2025. A survey on diffusion policy for robotic manipulation: Taxonomy, analysis, and future directions. *Authorea Preprints*.

Stepputtis, S.; Bandari, M.; Schaal, S.; and Amor, H. B. 2022. A system for imitation learning of contact-rich bimanual manipulation policies. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11810–11817. IEEE.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 5026–5033. IEEE.

Tsuji, T.; Kato, Y.; Solak, G.; Zhang, H.; Petrič, T.; Nori, F.; and Ajoudani, A. 2025. A Survey on Imitation Learning for Contact-Rich Tasks in Robotics.

Urain, J.; Mandlekar, A.; Du, Y.; Shafiullah, M.; Xu, D.; Fragkiadaki, K.; Chalvatzaki, G.; and Peters, J. 2024. Deep Generative Models in Robotics: A Survey on Learning from Multimodal Demonstrations. *CoRR*.

Wang, C.; Luo, X.; Ross, K.; and Li, D. 2022. Vrl3: A data-driven framework for visual deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 32974–32988.

Wolf, R.; Shi, Y.; Liu, S.; and Rayyes, R. 2025. Diffusion Models for Robotic Manipulation: A Survey.

Wu, Y.; Chen, Z.; Wu, F.; Chen, L.; Zhang, L.; Bing, Z.; Swikir, A.; Haddadin, S.; and Knoll, A. 2024. Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation.

Xie, Z.; Clary, P.; Dao, J.; Morais, P.; Hurst, J.; and Panne, M. 2020. Learning locomotion skills for cassie: Iterative design and sim-to-real. In *Conference on Robot Learning*, 317–329. PMLR.

Xue, H.; Ren, J.; Chen, W.; Zhang, G.; Fang, Y.; Gu, G.; Xu, H.; and Lu, C. 2025. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation.

Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; and Levine, S. 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, 1094–1100. PMLR.

Ze, Y.; Zhang, G.; Zhang, K.; Hu, C.; Wang, M.; and Xu, H. 2024. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. In *Proceedings of Robotics: Science and Systems (RSS)*.

Zhang, Q.; Liu, Z.; Fan, H.; Liu, G.; Zeng, B.; and Liu, S. 2025a. Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14754–14762.

Zhang, Z.; Chen, R.; Yang, Z.; Xie, S.; Chen, H.; and Xiong, H. 2025b. Unifying Modern AI with Robotics: Survey on MDPs with Diffusion and Foundation Models. *Authorea Preprints*.

Zhao, J.; Kuppuswamy, N.; Feng, S.; Burchfiel, B.; and Adelson, E. 2025. PolyTouch: A Robust Multi-Modal Tactile Sensor for Contact-rich Manipulation Using Tactile-Diffusion Policies.

Zhao, Z.; He, Y.; Yu, F.; Li, P.; Zhuo, F.; and Sun, X. 2024. LLaKey: Follow My Basic Action Instructions to Your Next Key State. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9604–9611. IEEE.