frontiers | Frontiers in Robotics and AI

TYPE Review
PUBLISHED 09 September 2025
DOI 10.3389/frobt.2025.1606247

OPEN ACCESS

EDITED BY
David Howard,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

REVIEWED BY
Fan Zhang,
Imperial College London, United Kingdom
Ananth Jonnavittula,
Virginia Tech, United States

*CORRESPONDENCE
Rosa Wolf,
rosa.wolf@kit.edu

RECEIVED 04 April 2025
ACCEPTED 14 July 2025
PUBLISHED 09 September 2025

CITATION
Wolf R, Shi Y, Liu S and Rayyes R (2025) Diffusion models for robotic manipulation: a survey.
Front. Robot. AI 12:1606247.
doi: 10.3389/frobt.2025.1606247

Diffusion models for robotic manipulation: a survey

Rosa Wolf*, Yitian Shi, Sheng Liu and Rania Rayyes

AI and Robotics (AIR), Institute of Material Handling and Logistics (IFL), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Diffusion generative models have demonstrated remarkable success in visual domains such as image and video generation. They have also recently emerged as a promising approach in robotics, especially in robot manipulations. Diffusion models leverage a probabilistic framework, and they stand out with their ability to model multi-modal distributions and their robustness to high-dimensional input and output spaces. This survey provides a comprehensive review of state-of-the-art diffusion models in robotic manipulation, including grasp learning, trajectory planning, and data augmentation. Diffusion models for scene and image augmentation lie at the intersection of robotics and computer vision for vision-based tasks to enhance generalizability and data scarcity. This paper also presents the two main frameworks of diffusion models and their integration with imitation learning and reinforcement learning. In addition, it discusses the common architectures and benchmarks and points out the challenges and advantages of current state-of-the-art diffusion-based methods.

KEYWORDS

diffusion models, robot manipulation learning, generative models, imitation learning, grasp learning

1 Introduction

Diffusion Models (DMs) have emerged as highly promising deep generative models in diverse domains, including computer vision (Ho et al., 2020; Song J. et al., 2021; Nichol and Dhariwal, 2021; Ramesh et al., 2022; Rombach et al., 2022a), natural language processing (Li et al., 2022; Zhang et al., 2023; Yu et al., 2022), and robotics (Chi et al., 2023; Urain et al., 2023). DMs intrinsically posses the ability to model any distribution. They have demonstrated remarkable performance and stability in modeling complex and multi-modal distributions[1] from high-dimensional and visual data surpassing the ability of Gaussian Mixture Models (GMMs) or Energy-based models (EBMs) like Implicit behavior cloning (IBC) (Chi et al., 2023). While GMMs and IBCs can model multi-modal distributions, and IBCs can even learn complex discontinuous distributions (Florence et al., 2022), experiments (Chi et al., 2023) show that in practice, they might be heavily biased toward specific modes. In general, DMs have also demonstrated performance exceeding generative adversarial networks (GANs) (Krichen, 2023), which were previously considered the leading

1 In the context of probability distributions, "multi-modal" does not refer to multiple input modalities but rather to the presence of multiple peaks (modes) in the distribution, each representing a distinct possible outcome. For example, in trajectory planning, a multi-modal distribution can capture multiple feasible trajectories. Accurately modeling all modes is crucial for policies, as it enables better generalization to diverse scenarios during inference

paradigm in the field of generative models. GANs usually require adversarial training, which can lead to mode collapse and training instability (Krichen, 2023). Additionally, GANs have been reported to be sensitive to hyperparameters (Lucic et al., 2018).

Since 2022, there has been a noticeable increase in the implementation of diffusion probabilistic models within the field of robotic manipulation. These models are applied across various tasks, including trajectory planning, e.g. (Chi et al., 2023), and grasp prediction, e.g., (Urain et al., 2023). The ability of DMs to model multi-modal distributions is a great advantage in many robotic manipulation applications. In various manipulation tasks, such as trajectory planning and grasping, there exist multiple equally valid solutions (redundant solutions). Capturing all solutions improves generalizability and robots' versatility, as it enables generating feasible solutions under different conditions, such as different placements of objects or different constraints during inference. Although in the context of trajectory planning using DMs, primarily imitation learning is applied, DMs have been adapted for integration with reinforcement learning (RL), e.g., (Geng et al., 2023). Research efforts focus on various components of the diffusion process adapted to different tasks in the domain of robotic manipulation. To give just some examples, developed architectures integrate different or even multiple input modalities. One example of an input modality could be point clouds (Ze et al., 2024; Ke et al., 2024). With the provided depth information, models can learn more complex tasks, for which a better 3D scene understanding is crucial. Another example of an additional input modality could be natural language (Ke et al., 2024; Du et al., 2023; Li et al., 2025), which also enables the integration of foundation models, like large language models, into the workflow. In Ze et al. (2024), both point clouds and language task instructions are used as multiple input modalities. Others integrate DMs into hierarchical planning (Ma X. et al., 2024; Du et al., 2023) or skill learning (Liang et al., 2024; Mishra et al., 2023), to facilitate their state-of-the-art capabilities in modeling high-dimensional data and multi-modal distributions, for long-horizon and multi-task settings. Many methodologies, e.g., (Kasahara et al., 2024; Chen Z. et al., 2023), employ diffusion-based data augmentation in vision-based manipulation tasks to scale up datasets and reconstruct scenes. It is important to note that one of the major challenges of DMs is its comparatively slow sampling process, which has been addressed in many methods, e.g., (Song J. et al., 2021; Chen K. et al., 2024; Zhou H. et al., 2024), also enabling real-time prediction.

To the best of our knowledge, we provide the first survey of DMs concentrating on the field of robotic manipulation. The survey offers a systematic classification of various methodologies related to DMs within the realm of robotic manipulation, regarding network architecture, learning framework, application, and evaluation. Alongside comprehensive descriptions, we present illustrative taxonomies.

To provide the reader with the necessary background information on DMs, we will first introduce their fundamental mathematical concepts (Section 2). This section provides a general overview of DMs rather than focusing specifically on robotic manipulation. Then, network architectures commonly used for DMs in robotic manipulation will be discussed (Section 3). Next (Section 4), we explore the three primary applications of DMs in robotic manipulation: trajectory generation (Section 4.1), robotic grasp synthesis (Section 4.2), and visual data augmentation (Section 4.3). This is followed by an overview of commonly used benchmarks and baselines (Section 5). Finally, we discuss our conclusions and existing limitations, and outline potential directions for future research (Section 6).

2 Preliminaries on diffusion models

2.1 Mathematical framework

The key idea of DMs is to gradually perturb an unknown target distribution $p_{\text{data}}(x)$ into a simple known distribution, e.g., a normal Gaussian distribution, which is first introduced in (Sohl-Dickstein et al., 2015). To generate new data, points are sampled from the initial known "simple" distribution, and perturbations are estimated to iteratively reverse the diffusion process. The forward and backward diffusion processes are also visualized in Figure 1. There exist two main approaches to diffusion-based modeling, both based on the original work by Sohl-Dickstein et al. (2015). The first group of methods is score-based DMs, where the gradient of the log-likelihood of the data is learned to reverse the diffusion process. This score-based generative modeling was first introduced in Song and Ermon (2019). In the other group of methods, a network is trained to directly predict the noise, which is added during the forward process. This methodology was first introduced in Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020).

The original score-based DM by Song and Ermon (2019) is rarely used in the field of robotic manipulation. This could be due to its inefficient sampling process. However, as it forms a crucial mathematical framework and baseline for many of the later developed DMs, e.g., (Song Y. et al., 2021; Karras et al., 2022), including DDPM Ho et al. (2020), we describe the main concepts in the following section. While DDPM is rarely used as well, the commonly used method Denoising Diffusion Implicit Models (DDIM) (Song J. et al., 2021) originates from DDPM. DDIM only alters the sampling process of DDPM while keeping its training procedure. Hence, understanding DDPM is crucial for many applications of DMs in robotic manipulation.

In the following sections, we first introduce score-based DMs, then DDPM, before addressing their shortcomings.

2.1.1 Denoising score matching using Noise Conditional Score Networks

One approach to estimate perturbations in the data distribution is to use denoising score matching with Lagenvin dynamics (SMLD), where the score of the data density of the perturbed distributions is learned using a Noise Conditional Score Network (NCSM) (Song and Ermon, 2019). This method is described in this section, and for more details, please refer to their original work. During the forward diffusion process, data $\mathbf{x}$ from an unknown distribution $p_{\text{data}}(\mathbf{x})$ is transformed into random noise $\mathcal{N}(0,\mathrm{I})$, by gradually adding noise. New data is generated during the reverse process, where the learned NCSM is used to iteratively denoise the initial samples.

2.1.1.1 Forward process

Let $\{\sigma_k\}_{k=1}^{K}$ be a noise schedule with progressively increasing variance, i.e., $\sigma_k < \sigma_{k+1}$ for all $k \in \{1,\ldots,K\}$. To get from the true data

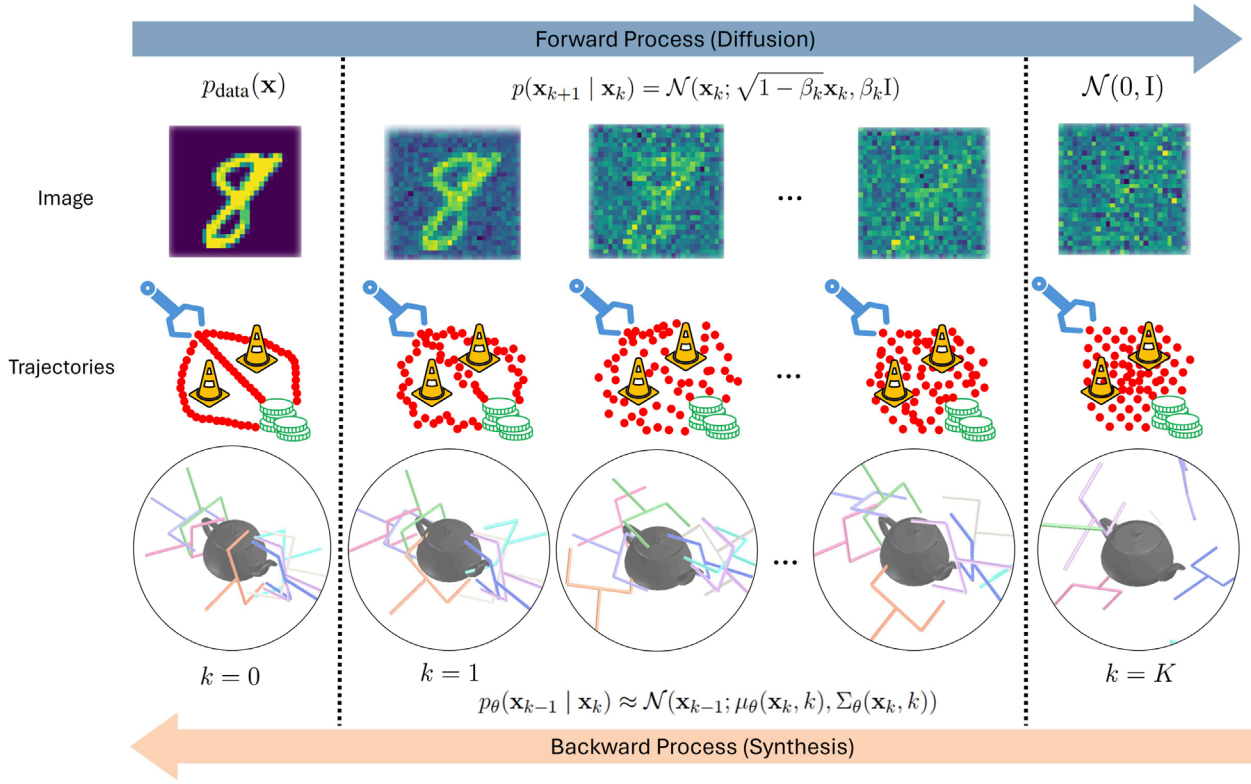

FIGURE 1
Illustrations of diffusion (forward) processes on image, trajectories, and grasp poses (Urain et al., 2023) and their corresponding synthesis (backward) processes.

distribution $p_{\text{data}}(\mathbf{x})$ to the perturbed data distribution $p_{\sigma_k}(\mathbf{x}_k)$, with variance σ_k, noise is added to the data according to a pre-specified noise distribution $p_{\sigma_k}(\mathbf{x}_k \mid \mathbf{x})$. To denoise the data, the gradients of the logarithmic probability density functions $\nabla_{\mathbf{x}} \log p_{\sigma_k}(\mathbf{x}_k \mid \mathbf{x})$, i.e., the scores, are estimated using the NCSM. To train the NCSM $\mathbf{s}_\theta(\mathbf{x}_k, \sigma_k)$, for all noise scales $k \in \{1, \ldots, K\}$ the weighted sum of denoising score matchings is minimized (Song and Ermon, 2019):

$$\mathcal{L} = \frac{1}{2K} \sum_{k=1}^{K} \sigma_k^2 \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{\mathbf{x}_k \sim p_{\sigma_k}(\mathbf{x}_k \mid \mathbf{x})} \left[\left\| \nabla_{\mathbf{x}_k} p_{\sigma_k}(\mathbf{x}_k \mid \mathbf{x}) - \mathbf{s}_\theta(\mathbf{x}_k, \sigma_k) \right\|_2^2 \right]. \tag{1}$$

2.1.1.2 Reverse process

Starting with randomly drawn noise samples $x_K^0 \in \mathcal{N}(0, I)$, Langevin dynamics are applied recursively over all $k \in \{0, \ldots, K\}$, to generate samples using the learned score function:

$$\mathbf{x}_k^n = \mathbf{x}_k^{n-1} + \alpha_k \mathbf{s}_\theta\left(x_k^{n-1}, \sigma_k\right) + \sqrt{2\alpha_k}\mathbf{z}_k^n, \quad n \in \{0, ., N\}, \tag{2}$$

where $\alpha_k > 0$ is the step size and $z_k^n \in \mathcal{N}(0, I)$ is randomly drawn noise. During one Langevin dynamic for noise scale k, the index n is increasing until $n = N$. Then, the final value $\mathbf{x}_k^N$, of one Langevin dynamic becomes the initial value $\mathbf{x}_{k-1}^0$ for the next Langevin dynamic with the next lower noise scale $k-1$, i.e., $x_{k-1}^0 = x_k^N$. For small enough step sizes, the final generated samples $\mathbf{x}_0^N$, should be approximately distributed according to $p_{\text{data}}(\mathbf{x})$.

2.1.2 Denoising Diffusion Probabilistic Models (DDPM)

In DDPM (Ho et al., 2020), instead of estimating the score function directly, a noise prediction network, conditioned on the noise scale, is trained. Similarly to SMLD with NCSN, new points are generated by sampling Gaussian noise and iteratively denoising the samples using the learned noise prediction network.

Notably, there is one step per noise scale in the denoising process instead of recursively sampling from each noise scale.

2.1.2.1 Forward process

To train the noise prediction network ϵ_θ, first points $\mathbf{x_0} \sim p_{\text{data}}(\mathbf{x})$ are sampled from the true unknown data distribution. The samples are degraded by adding noise $\epsilon \in \mathcal{N}(0, \mathrm{I})$ until at degrading step K, the degraded samples are approximately normally distributed, i.e. $x_K \sim \mathcal{N}(0, \mathrm{I})$. As already introduced by Sohl-Dickstein et al. (2015), the noise is added according to a Markovian process:

$$p\left(\mathbf{x}_{k+1} \mid \mathbf{x}_k\right) = \mathcal{N}\left(\mathbf{x}_k; \sqrt{1-\beta_k}\mathbf{x}_k, \beta_k \mathrm{I}\right), \tag{3}$$

where $\beta_1, \ldots, \beta_K \in [0, 1)$ is the noise variance schedule, which can either be a hyperparameter (Ho et al., 2020), or optimized as part of the model training process (Nichol and Dhariwal, 2021). In practice, instead of adding noise iteratively, the formulation also allows adding the noise in closed form:

$$p\left(\mathbf{x}_{k+1} \mid \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_k; \sqrt{\bar{\alpha}_k}\mathbf{x}_0, (1-\bar{\alpha}_k)\mathrm{I}\right), \tag{4}$$

with $\bar{\alpha}_k \coloneqq \prod_{i=1}^{k}(\alpha_i)$ and $\alpha_k \coloneqq 1-\beta_k$. This allows first uniformly sampling a noise scale $k \sim \mathcal{U}\{1,K\}$, and then directly inferring the corresponding degraded sample.

Adding the noise in closed form facilitates training a noise prediction network $\epsilon_\theta(\mathbf{x}_k,k)$ by minimizing the mean squared error for $k \in \{1,\dots,K\}$:

$$\mathcal{L} = \mathbb{E}_{k,\mathbf{x}_0,\epsilon}\left[||\epsilon - \epsilon_\theta(\mathbf{x}_k,k)||_2^2\right]. \tag{5}$$

2.1.2.2 Reverse process

Similar to the reverse process described in Section 2.1.1, new samples are generated from random noise $\mathbf{x}_K \sim \mathcal{N}(0,\mathrm{I})$, using the learned forward process $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$. As the forward process is modeled using Gaussian distributions, the reverse process $p_\theta(\mathbf{x}_{k-1} \mid \mathbf{x}_k)$ is also a Gaussian distribution if the number of diffusion steps is sufficiently large, i.e., the step size is small enough (Sohl-Dickstein et al., 2015):

$$p_\theta(\mathbf{x}_{k-1} \mid \mathbf{x}_k) \approx \mathcal{N}\left(\mathbf{x}_{k-1}; \mu_\theta(\mathbf{x}_k,k), \Sigma_\theta(\mathbf{x}_k,k)\right). \tag{6}$$

In DDPM, the variance-schedule is fixed and thus $\Sigma_\theta(x_k,k) = \beta_k \mathrm{I}$. Additionally, using reparameterization, it can be shown that the mean of the distribution at each step can be iteratively predicted using the previous value $\mathbf{x}_k$ and the estimated noise ϵ_θ (Ho et al., 2020):

$$\mathbf{x}_{k-1} = \frac{1}{\sqrt{\alpha_k}}\left(\mathbf{x}_k - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(\mathbf{x}_k,k)\right) + \sigma_k \mathbf{z}, \tag{7}$$

which is repeated until $\mathbf{x}_0$ is computed. As in SMLD, for small enough step sizes, the final generated samples $\mathbf{x}_0$ are approximately distributed according to the true data distribution $p_{\text{data}}(x)$.

2.2 Architectural improvements and adaptations

One of the main disadvantages of DMs is the iterative sampling, leading to a relatively slow sampling process. In comparison, using GANs or variational autoencoders (VAEs), only a single forward pass through the trained network is required to produce a sample. In both DDPM and the original formulation of SMLD, the number of time steps (noise levels) in the forward and reverse processes is equal. While reducing the number of noise levels leads to a faster sampling process, it comes at the cost of sample quality. Thus, there have been numerous works to adapt the architectures and sampling processes of DDPM and SMLD to improve both the sampling speed and quality of DMs, e.g., (Nichol and Dhariwal, 2021; Song J. et al., 2021; Song Y. et al., 2021).

2.2.1 Improving sampling speed and quality

The forward diffusion process can be formulated as a stochastic differential equation (SDE). Using the corresponding reverse-time SDE, SDE-solvers can then be applied to generate new samples (Song Y. et al., 2021). Song et al. (2021b) shows that the diffusion process from SMLD corresponds to an SDE where the variance of the perturbation kernels $\{p(x_k \mid x_0)\}_{k=1}^{K}$ is exploding with increasing K. This is referred to as the variance exploding SDE (VE SDE) in the literature. The diffusion process from DDPM corresponds to a variance-preserving SDE, referred to as VP SDE in the literature. As such, the original formulations of SMLD and DDPM can be interpreted as specific discretizations of their corresponding SDEs. Song Y. et al. (2021) also shows that once the score-network is trained, the reverse-time SDE can be replaced by an ordinary differential equation (ODE). Using an ODE has several advantages. As the reverse process is deterministic, it allows for precise likelihood computation (Song Y. et al., 2021). Moreover, the deterministic process naturally leads to higher consistency. Thus, the ODE formulation can be used as a high-level feature-preserving encoding, which also allows interpolations in latent space (Song J. et al., 2021; Karras et al., 2022). Finally, using ODEs enables faster and adaptive sampling, which is why it forms the baseline for many of the following methods.

One group of methods aimed at improving sampling speed (Jolicoeur-Martineau et al., 2021; Song J. et al., 2021; Lu et al., 2022; Karras et al., 2022) designs samplers that operate independently of the specific training process. Using an SDE/ODE-based formulation allows choosing different discretizations of the reverse process than for the forward process. Larger step sizes reduce computational cost and sampling time but introduce greater truncation error. The sampler operates independently of the specific noise prediction network implementation, enabling the use of a single network, such as one trained with DDPM, with different samplers.

Denoising Diffusion Implicit Models (DDIM) (Nichol and Dhariwal, 2021) is the dominant method used for robotic manipulation. It uses a deterministic sampling process and outperforms DDPM when using only a few (10–100) sampling iterations. DDIM can be formulated as a first-order ODE solver. In Diffusion Probabilistic Models-solver (DPM-solver) (Lu et al., 2022), a second-order ODE solver is applied, which decreases the truncation error, thus further increasing performance on several image classification benchmarks for a low number of sampling steps. In contrast to DDIM, Karras et al. (2022); Lu et al. (2022) use non-uniform step sizes in the solver. In a detailed analysis Karras et al. (2022) empirically shows that compared to uniform step-sizes, linear decreasing step sizes during denoising lead to increased performance (Karras et al., 2022), indicating that errors near the true distribution have a larger impact.

Even though DPM-solver (Lu et al., 2022) shows superior performance over DDIM. It should be noted that in the original papers (Song J. et al., 2021; Lu et al., 2022), only image-classification benchmarks are considered to compare both methods. Therefore, more extensive tests should be performed to validate these results.

A second group of methods addressing sampling speed also adapts the training process or requires additional fine-tuning. Examples are knowledge distillation of DMs to gradually reduce the number of noise levels (Salimans and Ho, 2022), or finetuning of the noise schedule (Nichol and Dhariwal, 2021; Watson et al., 2022). While in DDPM and DDIM, the noise schedule is fixed, in improved Denoising Diffusion Probabilistic Models (iDDPM) (Nichol and Dhariwal, 2021), the noise schedule is learned, resulting in better sample quality. They also suggest changing from a linear noise schedule, like in DDPM, to other schedules, e.g., a cosine noise schedule. In particular, for low-resolution samples, a linear schedule leads to a noisy diffusion process with too rapid information

loss, while the cosine noise schedule has smaller steps during the beginning and end of the diffusion process. Already after a fraction of around 0.6 diffusion steps, the linear noise schedule is close to zero (and the data distribution close to white noise). Thus, the first steps of the reverse process do not strongly contribute to the data generation process, making the sampling process inefficient. Although iDDPM (Nichol and Dhariwal, 2021) also outperforms DDIM, it requires fine-tuning, which might be a reason why it is less popular.

There are also several methods (Zhou H. et al., 2024; Li X. et al., 2024; Wang et al., 2023b; Chen K. et al., 2024) regarding sampling speed, specifically for applications in robotic manipulation, which is different from the previously named methodologies, which were developed in the context of image processing. For example, Chen K. et al. (2024) samples from a more informed distribution than a Gaussian. They point out that even initial distributions approximated with simple heuristics result in better sample quality, especially when using few diffusion steps or when only a limited amount of data is available. Others (Prasad et al., 2024) use teacher–student distillation techniques (Tarvainen and Valpola, 2017), where pretrained diffusion models serve as teachers, guiding student models to operate with larger denoising steps while preserving consistency with the teacher's results at smaller steps. While this increases training effort, it decreases sampling time at inference, which is especially important in (near) real-time control.

Recently, flow matching (Lipman et al., 2023) has been used as an alternative method to diffusion. Like with diffusion, the true distribution is estimated starting from a noise distribution. However, instead of learning the time-dependent score or noise, and then deriving the velocity from noise to data distribution from it, in flow matching, the time-dependent velocity field is learned directly. This leads to a simpler training objective, using the interpolation between the noise sample and true data point, without requiring a noise schedule. Thus, flow matching is usually more numerically stable and requires less hyperparameter tuning. However, when using few sampling steps, with flow matching, there is a risk of mode-collapse and infeasible solutions, as the ODE-solver averages over the velocity field. Thus, Frans et al. (2025) conditions the model not only on the time-step, but also on the step-size. By using the fact that one large step should lead to the same point as two consecutive steps of half the size, they maximize a self-consistency objective in addition to the flow-matching objective. Thus, the model can sample with a single step, with only a small drop in performance, far surpassing the performance of DDIM, when only a small number of sampling steps are used. While this is similar to the above-mentioned distillation techniques (Prasad et al., 2024), here only a single model has to be trained.

2.3 Adaptations for robotic manipulation

Two main points must be considered to apply DMs to robotic manipulation. Firstly, in the diffusion processes described in the previous sections, given the initial noise, samples are generated solely based on the trained noise prediction network or conditional score network. However, robot actions are usually dependent on simulated or real-world observations with multi-modal sensory data and the robot's proprioception. Thus, the network used in the denoising process has to be conditioned on these observations (Chi et al., 2023). Encoding observations varies in different algorithms. Some use ground truth state information, such as object positions (Ada et al., 2024), and object features, like object sizes (Mishra et al., 2023; Mendez-Mendez et al., 2023). In this case, sim-to-real transfer is challenging due to sensor inaccuracies, object occlusions, or other adversarial settings, e.g., lightning conditions, Therefore, most methods directly condition on visual observations, such as images (Si et al., 2024; Bharadhwaj et al., 2024a; Vosylius et al., 2024; Chi et al., 2023; Shi et al., 2023), point clouds (Liu et al., 2023c; Li et al., 2025), or feature encodings and embeddings (Ze et al., 2024; Ke et al., 2024; Li X. et al., 2024; Pearce et al., 2022; Liang et al., 2024; Xian et al., 2023; Xu et al., 2023), where the robustness to adversarial setting can be directly addressed.

Secondly, unlike in image generation, where the pixels are spatially correlated, in trajectory generation for robotic manipulation, the samples of a trajectory are temporally correlated. On the one hand, generating complete trajectories may not only lead to high inaccuracies and error accumulation of the long-horizon predictions, but also prevent the model from reacting to changes in the environment. On the other hand, predicting the trajectory one action at a time increases the compounding error effect and may lead to frequent switches between modes. Accordingly, trajectories are mostly predicted in subsequences, with a receding horizon, e.g., (Chi et al., 2023; Scheikl et al., 2024), which will be discussed in more detail in Section 4.1 and is visualized in Figure 2. In receding horizon control, the diffusion model generates only a subtrajectory with each backward pass. The subtrajectory is executed before generating the next subtrajectory on the updated observations. In comparison, grasps are generated similarly to images. As here only a single action, usually the grasp pose, is generated, this is done using a single backward pass of the diffusion model. Moreover, the grasp pose is usually predicted from a single initial observation. During execution, possible changes in the scene are not being taken into account. The backward pass for generating one action is visualized in Figure 1.

3 Architecture

3.1 Network architecture

For the implementation of the DM, it is essential to select an appropriate architecture for the noise prediction network. There exist three predominant architectures used for the denoising diffusion networks: Convolutional neural networks (CNNs), transformers, and Multi-Layer Perceptrons (MLPs).

3.1.1 Convolutional neural networks

The most frequently employed architecture is the CNN, more specifically the Temporal U-Net that was first introduced by Janner et al. (2022) in their algorithm Diffuser, a DM for robotics tasks. The U-Net architecture (Ronneberger et al., 2015) has shown great success in image generation with DMs, e.g., (Ho et al., 2020; Dhariwal and Nichol, 2021; Song Y. et al., 2021). U-net, in general, is proven to be sample efficient and can even generalize well with small training datasets (Meyer-Veit et al., 2022b; Meyer-Veit et al., 2022a).

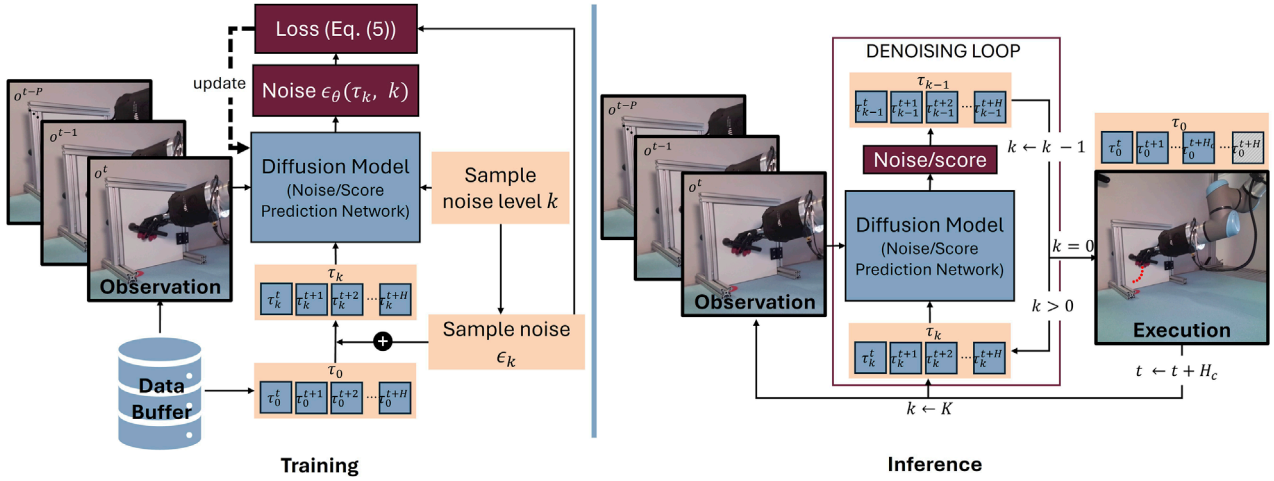

FIGURE 2
Illustrations of the iterative trajectory generation using receding horizon control. At inference, the trajectory is planned up to a planning horizon H, conditioned on the past P observations $\{o_t, o_{t-1}, \ldots, o_{t-P}\}$. Of this plan, only the steps until the control horizon $H_c \leq H$ are executed. In the figure, this is visualized in the outer loop with the time variable t. In the inner denoising loop, one subtrajectory $\tau = \{\tau_t, \tau_{t+1}, \ldots, \tau_{t+H}\}$ at the current time step t is generated, using a diffusion model. Conditioned on the last P observations and the current noise level k, the diffusion model predicts the noise, or score, dependent on the model type. Using the predicted noise/score, the trajectory at the next lower noise level $k-1$ is calculated. This is then used as the next input to the diffusion model until the trajectory is completely denoised ($k = 0$), at which point it is executed. After execution of the subtrajectory, the time is increased and the next H steps of the trajectory are planned. For training, ground truth trajectories and corresponding observations are sampled from the data buffer. The diffusion model is also trained on subtrajectories. However, the lookahead H during training may be chosen larger than during inference, to ensure flexibility. The diffusion model is trained to predict the noise of a noisy trajectory. For this, first, a noise level k is sampled. Then the noise ϵ_k is sampled, according to the predefined variance schedule. The noise is added in closed form to the ground-truth trajectory τ_0 (see Equation 4) to get the noisy trajectory τ_k. The predicted noise $\epsilon_\theta(\tau_k, k)$ on the trajectory τ_k is compared with the true sampled noise ϵ_k to compute the loss. Using this, the diffusion model can be updated.

Thus, it has been adapted to robotic manipulation by replacing two-dimensional spatial convolutions with one-dimensional temporal convolutions (Janner et al., 2022).

The temporal U-Net is further adapted by Chi et al. (2023) in their CNN-based Diffusion Policy (DP) for robotic manipulation. While in Diffuser, the state and action trajectories are jointly denoised, only the action trajectories are generated in DP. To ensure temporal consistency, the diffusion process is conditioned on a history of observations using feature-wise linear modification (FiLM) (Perez et al., 2018). This formulation allows for an extension to different and multiple conditions by concatenating them in feature space before applying FiLM (Li X. et al., 2024; Si et al., 2024; Ze et al., 2024; Li et al., 2025; Wang L. et al., 2024). Moreover, it also enables the incorporation of constraints embedded with an MLP (Ajay et al., 2023; Zhou et al., 2023; Power et al., 2023).

Discussed in more detail in Section 4.1.1.6, Janner et al. (2022) formulates conditioning as inpainting, where during inferences at each denoising step, specific states from the currently being generated sample are replaced with states from the condition. For example, the final state of a generated trajectory may be replaced by the goal state, for goal-conditioning. This only affects the sampling process at inference and, thus, does not require any adaptations of the network architecture. However, it only supports point-wise conditions, severely limiting its applications. Multiple frameworks (Saha et al., 2024; Carvalho et al., 2023; Wang et al., 2023b; Ma X. et al., 2024) directly employ the temporal U-Net architecture introduced by Janner et al. (2022). However, as this type of conditioning is highly limited in its applications, FiLM conditioning is more common. A different but less-used architecture incorporates conditions via cross-attention mapped to the intermediate layers of the U-Net (Zhang E. et al., 2024), which is more complicated to integrate than FiLM conditioning.

3.1.2 Transformers

Another commonly used architecture for the denoising network are transformers. A history of observations, the current denoising time step, and the (partially denoised) action are input tokens to the transformer. Additional conditions can be integrated via self-and cross-attention, e.g., (Chi et al., 2023; Mishra and Chen, 2024). The exact architecture of the transformer varies across methods. The more commonly used model is a multi-head cross-attention transformer as the denoising network, e.g., (Chi et al., 2023; Pearce et al., 2022; Wang et al., 2023b; Mishra and Chen, 2024). Others (Bharadhwaj et al., 2024b; Mishra et al., 2023) use architectures based on the method Diffusion Transformers (Peebles and Xie, 2023), which is the first method combining DMs with transformer architectures. There are also less commonly used architectures, such as using the output tokens of the transformer as input to an MLP, which predicts the noise (Ke et al., 2024).

For completeness, we provide a list of works, using transformer architectures: (Chi et al., 2023; Pearce et al., 2022; Scheikl et al., 2024; Wang et al., 2023b; Ze et al., 2024; Feng et al., 2024; Bharadhwaj et al., 2024b; Mishra et al., 2023; Liu et al., 2023b; Xu et al., 2024; Mishra and Chen, 2024; Liu et al., 2023c; Vosylius et al., 2024; Reuss et al., 2023; Iioka et al., 2023; Huang T. et al., 2025).

3.1.3 Multi-Layer Perceptrons

Predominantly used for applications in RL, MLPs are employed as denoising networks, e.g., (Suh et al., 2023; Ding and Jin, 2023; Pearce et al., 2022), which take concatenated input features, such as observations, actions, and denoising time steps, to predict the noise. Although the architectures vary, it is common to use a relatively small number of hidden layers (2–4) (Wang et al., 2023b; Kang et al., 2023; Suh et al., 2023; Mendez-Mendez et al., 2023), using e.g., Mish activation (Misra, 2019), following the first method (Wang et al., 2023a), integrating DMs with Q-learning. It is important to note that most of these methods do not use visual input. An exception from this is Pearce et al. (2022), which also evaluates using high-resolution image inputs with an MLP-based DM. However, for this, a CNN-based image encoder is first applied to the raw image observation, before the encoding is fed to the DM.

3.1.4 Comparison

An ongoing debate exists concerning the relative merits of different architectural choices, with each architecture exhibiting distinct advantages and disadvantages. Chi et al. (2023) implemented both a U-Net-based and a transformer-based denoising network with the application of trajectory planning. They observed that the CNN-based model exhibits lower sensitivity to hyperparameters than transformers. Moreover, they report that when using positional control, the U-net results in a slightly higher success rate for some complex visual tasks, such as transport, tool hand, and push-t. On the other hand, U-nets may induce an over-smoothing effect, thereby resulting in diminished performance for high-frequency trajectories and consequently affecting velocity control. Thus, in these cases, transformers will likely lead to more precise predictions. Furthermore, transformer-based architectures have demonstrated proficiency in capturing long-range dependencies and exhibit notable robustness when handling high-dimensional data, surpassing the abilities of CNNs, which is particularly significant for tasks involving long horizons and high-level decision-making (Janner et al., 2022; Dosovitskiy et al., 2021).

While MLPs typically exhibit inferior performance, especially when confronted with complex problems and high-dimensional input data, such as images, they often demonstrate superior computational efficiency, which facilitates higher-rate sampling and usually requires fewer computational resources. Due to their training stability, they are a commonly used architecture in RL. In contrast, U-Nets, and especially transformers, are characterized by substantial resource consumption and prolonged inference times, which may hinder their application in real-time robotics (Pearce et al., 2022).

In summary, transformers are the most powerful architecture for handling high-dimensional input and output spaces, followed by CNNs, while MLPs have the highest computational efficiency. For processing visual data, such as raw images, an important task in robotic manipulation, a CNN or a Transformer architecture should be chosen. Also, while MLPs are most computationally efficient, real-time control is possible with the other two architectures, integrating, for example, receding horizon control (Mattingley et al., 2011) in combination with a more efficient sampling process, like DDIM.

3.2 Number of sampling steps

In addition to the network architecture, a crucial decision is the choice of the number of training and sampling iterations. As described in Section 2.2, each sample must undergo iterative denoising over several steps, which can be notably time-consuming, especially in the context of employing larger denoising networks with longer inference durations, such as transformers. Within the framework of DDPM, the number of noise levels during training is equal to the number of denoising iterations at the time of inference. This hinders its use in many robotic manipulation scenarios, especially those necessitating real-time predictions. Consequently, numerous methodologies employ DDIM, where the number of sampling iterations during inference can be significantly reduced compared to the number of noise levels used during training. Common choices of noise levels are 50–100 during training, but only a subset of five to ten steps during inference (Chi et al., 2023; Ma X. et al., 2024; Huang T. et al., 2025; Scheikl et al., 2024). Only a few works used less sampling (3–4) (Vosylius et al., 2024; Reuss et al., 2023) or more (20–30) (Mishra and Chen, 2024; Wang L. et al., 2024) sampling steps. Ko et al. (2024) documented a slight decline in performance when the number of sampling steps is reduced to 10% with DDIM (Ko et al., 2024). Therefore, it is imperative to consider an appropriate trade-off between sample quality and inference time, tailored to the specific task requirements. Still, only a few evaluations exist that compare DDPM-based, DDIM-based, or other samplers for robotic manipulation, and further investigation is required.

4 Applications

In this section, we explore the most dominant applications of DMs in robotic manipulation: trajectory generation for robotic manipulation, robotic grasping, and visual data augmentation for vision-based robotics manipulations.

4.1 Trajectory generation

Trajectory planning in robotic manipulation is vital for enabling robots to move from one point to another smoothly, safely, and efficiently while adhering to physical constraints, like speed and acceleration limits, as well as ensuring collision avoidance. Classical planning methods, like interpolation-based and sampling-based approaches, can have difficulty handling complex tasks or ensuring smooth paths. For instance, Rapidly Exploring Random Trees (Martinez et al., 2023) might generate trajectories with sudden changes because of the discretization process. As already discussed in the introduction, although popular data-driven approaches, such as GMMs and EBMs, theoretically pertain to the ability to model multi-model data distributions, in reality, they show suboptimal behavior, such as biasing modes or lack of temporal consistency (Chi et al., 2023). In addition, GMMs can struggle with high-dimensional input spaces (Ho et al., 2020). Increasing the number of components and covariances also increases the models' ability to model more complex distributions and capture complex and intricate movement patterns. However, this can negatively impact

TABLE 1 Taxonomy of imitation learning approaches for trajectory generation with diffusion models.

Perspective	Category	Subcategory	References
Methodological	Actions and pose representations	Task Space §4.1.1.1	Chi et al. (2023), Pearce et al. (2022), Ze et al. (2024), Ha et al. (2023), Ke et al. (2024), Xu et al. (2023), Li et al. (2024c), Si et al. (2024), Scheikl et al. (2024), Xian et al. (2023), Liu et al. (2023c)
		Joint Space §4.1.1.1	Carvalho et al. (2023), Saha et al. (2024), Urain et al. (2023), Ma et al. (2024b)
		Image Space §4.1.1.3	Ko et al. (2024), Yang et al. (2024), Zhou et al. (2024b), Vosylius et al. (2024), Du et al. (2023), Liang et al. (2024)
	Visual data modality §4.1.1.2	2D	e.g. Chi et al. (2023), Liang et al. (2024), Scheikl et al. (2024), Si et al. (2024)
		3D	Li et al. (2025), Liu et al. (2023c), Wang et al. (2024a), Ze et al. (2024), Xian et al. (2023), Ke et al. (2024)
Functional	Long-Horizon and Multi-Task Learning	Hierarchical Planning §4.1.1.4	Zhang et al. (2024a), Ma et al. (2024b), Xian et al. (2023), Ha et al. (2023), Huang et al. (2024b), Du et al. (2023)
		Skill Learning §4.1.1.4	Mishra et al. (2023), Kim et al. (2024c), Xu et al. (2023), Liang et al. (2024)
		Vision Language Action Models §4.1.1.5	Pan et al. (2024a), Shentu et al. (2024), Team et al. (2024), Wen et al. (2025), Liu et al. (2024), Li et al. (2024b), Black et al. (2024a)
	Constrained Planning §4.1.1.6	Classifier guidance	Mishra et al. (2023), Liang et al. (2023), Janner et al. (2022), Carvalho et al. (2023)
		Classifier-free guidance	Ho et al. (2021), Saha et al. (2024), Li et al. (2025), Power et al. (2023), Reuss et al. (2024a), Reuss et al. (2023)

the smoothness of the generated trajectories, making GMMs highly sensitive to their hyperparameters. In contrast, denoising DMs have demonstrated exceptional performance in processing and generating high-dimensional data. Furthermore, the distributions generated by denoising DMs are inherently smooth (Ho et al., 2020; Sohl-Dickstein et al., 2015; Chi et al., 2023). This makes DMs well-suited for complex, high-dimensional scenarios where flexibility and adaptability are required. While most methodologies that apply probabilistic DMs to robotic manipulation focus on imitation learning, they have also been adapted to their application in RL, e.g., (Janner et al., 2022; Wang et al., 2023a).

In the following sections, the methodologies of DMs for trajectory generation will be further discussed and categorized. We will first explain their applications in imitation learning, followed by a discussion on their use in reinforcement learning. For an overview of the method architectures in imitation learning, see Table 2, and for reinforcement learning, see Table 3.

4.1.1 Imitation learning

In imitation learning (Zare et al., 2024), robots attempt to learn a specified task by observing multiple expert demonstrations. This paradigm, commonly known as Learning from Demonstrations (LfD), involves the robot observing expert examples and attempting to replicate the demonstrated behaviors. In this domain, the robot is expected to generalize beyond the specific demonstrations, which allows the robot to adapt to variations in tasks or changes in configuration spaces. This may include diverse observation perspectives, altered environmental conditions, or even new tasks that share structural similarities with those previously demonstrated. Thus, the robot must learn a representation of the task that allows flexibility and skill acquisition beyond the specific scenarios it was trained on. Recent advancements in applying DMs to learn visuomotor policies (Chi et al., 2023) enable the generation of smooth action trajectories by modeling the task as a generative process conditioned on sensory observations. Diffusion-based models, initially popularized for high-dimensional data generation such as images and natural languages, have demonstrated significant potential in robotics by effectively learning complex action distributions and generating multi-modal behaviors conditioned on task-specific inputs. For instance, combining with recent progress in multiview transformers (Gervet et al., 2023; Goyal et al., 2023) that leverage the foundation model features (Radford et al., 2021; Oquab et al., 2023), 3D diffuser actor (Ke et al., 2024) integrates multi-modal representations to generate the end-effector trajectories. As another example, GNFactor (Ze et al., 2023) renders multiview features from Stable Diffusion (Rombach et al., 2022b) to enhance 3d volumetric feature learning. Very similar to diffusion, recently (Rouxel et al., 2024) flow-matching-based policies have emerged for trajectory generation, generally leading to a more stable training process with fewer hyperparameters, as

TABLE 2 Technical details of trajectory diffusion using imitation learning. The references for the encoders are provided in Supplementary Appendix Table 1. In the following, the symbols and abbreviations are explained: H: Whether the method is hierarchical (✓) or not (✗). PCs: Point Clouds, Lan: Language, GTS: Ground Truth State, and whether the visual input modality is from single view or (^{SV}) multi-view (^{MV}). U-Net: temporal U-Net (Janner et al., 2022), FiLM: Convolutional Neural Networks with Feature-wise Linear Modulation (Perez et al., 2018), DiT: Diffusion Transformer, RHC: sub-trajectories with receding horizon control, CT: complete trajectory in task space, J: complete trajectory in joint space. A"/" indicates that the information is not provided by the cited paper, while a "-" indicates that no specialized encoder is required as ground truth state information is used.

Reference	Input	Output	Encoder	Diffuser	H
Chi et al. (2023)	RGB^{MV}	RHC	ResNet	FiLM	✗
Xian et al. (2023)	RGB-D ^{MV}, Lan	CT	CLIP	DiT & MLP	✓
Reuss et al. (2023)	GTS/RGB ^{SV}	CT	ResNet	DiT	✗
Chen et al. (2023a)	RGB ^{SV}, Lan	RHC	ResNet	U-Net	✗
Zhou et al. (2023)	RGB ^{MV}	RHC	CLIP	U-Net	✗
Pearce et al. (2022)	RGB ^{SV}	RHC	CNN/ResNet	MLP/DiT	✗
Mendez-Mendez et al. (2023)	GTS	RHC	-	MLPs	✓
Ze et al. (2024)	PCs ^{SV}	RHC	MLP	FiLM	✗
Ke et al. (2024)	RGB-D$^{SV/MV}$, Lan	CT	CLIP	DiT	✗
Power et al. (2023)	GTS	RHC	MLP	U-Net	✗
Ma et al. (2024b)	RGB-D ^{SV}, Lan	J	PointNet++, MLP	U-Net	✓
Vosylius et al. (2024)	RGB ^{MV}	RHC	Transformer	DiT	✗
Zhang et al. (2024a)	RGB^{SV}, Lan	RHC	HULC, T5	U-Net	✓
Reuss et al. (2024a)	RGB ^{MV}, Lan	RHC	ResNet, CLIP	DiT	✗
Scheikl et al. (2024)	RGB ^{SV}/GTS	RHC	ResNet	DiT	✗
Chen et al. (2024a)	GTS/PCs/RGB^{SV}	RHC	—	—	✗
Zhou et al. (2024a)	GTS/RGB^{SV}	RHC	ResNet	DiT	✓
Li et al. (2025)	PCs^{SV}, Lan	RHC	SAM, XMem	FiLM	✗
Li et al. (2024c)	RGB^{MV}	RHC	ResNet	FiLM	✗
Si et al. (2024)	RGB^{SV}	RHC	ResNet	FiLM	✗
Saha et al. (2024)	GTS	RHC	-	U-Net	✗
Bharadhwaj et al. (2024b)	RGB^{SV}	point tracks	—	DiT	✗
Wang et al. (2024b)	RGB, Tactile, PCs, Lan	RHC	ResNet, PointNet, T5	U-Net	✗
Reuss et al. (2024b)	RGB, Lan	RHC	ResNet, CLIP	DiT	✗

already mentioned in Section 2.2.1. Nguyen et al. (2025) additionally includes second-order dynamics into the flow-matching objective, learning fields on acceleration and jerk to ensure smoothness of the generated trajectories.

In terms of the type of robotic embodiment, most works use parallel grippers or simpler end-effectors. However, few methods perform dexterous manipulation using DMs (Si et al., 2024; Ma C. et al., 2024; Ze et al., 2024; Chen K. et al., 2024; Wang C. et al., 2024; Freiberg et al., 2025; Welte and Rayyes, 2025), to facilitate their stability and robustness, also in this high-dimensional setting.

In the following sections, we will first repeat the process of sampling actions for trajectory planning with DMs and discuss common pose representations. Then we shortly address different visual data modalities, in particular 2D vs. 3D visual observations. Afterwards, we look at methods formulating trajectory planning

TABLE 3 Technical details of trajectory diffusion using reinforcement learning. The references for the encoders are provided in Supplementary Appendix Table 1. In the following, the symbols and abbreviations are explained: H/S: Whether the method is hierarchical/skill-based (✓) or not (✗). Lan: Language, GTS: Ground Truth State, and whether the visual input modality is from single view (^{SV}) or multi-view (^{MV}). U-Net: temporal U-Net (Janner et al., 2022), Eq.: Equivariant FiLM: Convolutional Neural Networks with Feature-wise Linear Modulation (Perez et al., 2018), DiT: Diffusion Transformer, RHC: sub-trajectories with receding horizon control, Sia = single actions. A "-" indicates that no specialized encoder is required as ground truth state information is used.

Reference	Input	Output	Encoder	Diffuser	H/S
Janner et al. (2022)	GTS	RHC	-	U-Net	✗
Ajay et al. (2023)	GTS	RHC	-	U-Net	✓
Wang et al. (2023a)	GTS	SiA	-	MLP	✗
Wang et al. (2023b)	GTS	RHC	-	DiT	✗
Ding and Jin (2023)	GTS	SiA	-	MLP	✗
Mishra et al. (2023)	GTS	RHC	-	DiT	✓
Kang et al. (2023)	GTS	RHC	-	MLP	✗
Brehmer et al. (2023)	GTS	RHC	-	Eq. U-Net	✗
Suh et al. (2023)	GTS	RHC	-	U-Net	✗
Ha et al. (2023)	RGB^{MV}, Lan	RHC	ResNet, CLIP	FiLM	✓
Kim et al. (2024b)	GTS	RHC	-	U-Net	✓
Liang et al. (2023)	GTS	RHC	-	U-Net	✗
Ada et al. (2024)	GTS	SiA	-	MLP	✗
Ren et al. (2024)	RGB/GTS	SiA	ViT/-	U-Net/MLP	✗
Huang et al. (2025b)	RGB^{SV}	SiA	VQ-GAN	VQ-Diffusion Gu et al. (2022)	✗
Carvalho et al. (2023)	GTS	RHC	-	U-Net	✗

as image generation, before looking at applications in hierarchical, multi-task, and constrained planning, also looking at multi-task planning with vision language action models (VLAs). A visualization of the taxonomy is provided in Table 1. More details on the individual method architectures are provided in Table 2.

4.1.1.1 Actions and pose representation

As briefly discussed in Section 2.3, the entire trajectory can be generated as a single sample, multiple subsequences can be sampled using receding horizon control, or the trajectory can be generated by sampling individual steps. Only in a few methods (Janner et al., 2022; Ke et al., 2024) the whole trajectory is predicted at once. Although this enables a more efficient prediction, as the denoising has to be performed only once, it prohibits adapting to changes in the environment, requiring better foresight and making it unsuitable for more complex task settings with dynamic or open environments. On the other hand, sampling of individual steps increases the compounding error effect and can negatively affect temporal correlation. Instead of predicting micro-actions, some use DMs to predict waypoints (Shi et al., 2023). This can decrease the compounding error, by reducing the temporal horizon. However, it relies on preprocessing or task settings that ensure that the space in between waypoints is not occluded. Thus, typically, DMs generate trajectories consisting of sequences of micro-actions represented as end-effector positions, generally encompassing translation and rotation depending on end-effector actuation (Chi et al., 2023; Ze et al., 2024; Xu et al., 2023; Li X. et al., 2024; Si et al., 2024; Scheikl et al., 2024; Ke et al., 2024; Ha et al., 2023). Once the trajectory is sampled, the proximity of the predicted positions enables computing the motion between the positions with simple positional controllers without the need for complex trajectory planning techniques. The control scheme is visualized in detail in Figure 2. Although more commonly applied in grasp prediction, here the pose is sometimes also represented in special Euclidean group ($\mathbf{SE}(3)$) (Xian et al., 2023; Liu et al., 2023c; Ryu et al., 2024). Explained in more detail in Section 4.2, the group structure of the $\mathbf{SE}(3)$ Lie group enables continuous interpolation and transformations between multiple object poses. As Liu et al. (2023c), Ryu et al. (2024) performs complex tasks involving trajectory planning and grasping for aligning multiple objects, these properties are important to ensure physically and geometrically grounded actions. However, as the prediction of $\mathbf{SE}(3)$ poses with DMs requires a more complex model structure and training in imitation learning, it is more usual to use representations, such as Euler angles

or quaternions, in trajectory planning. Not only diffusion, but also flow matching has been adapted to use representations in **SE**(3) or Riemannian manifolds in general (Braun et al., 2024).

Although not common, sometimes actions are predicted directly in joint space (Carvalho et al., 2023; Pearce et al., 2022; Saha et al., 2024; Ma X. et al., 2024), allowing for direct control of joint motions, which, e.g., reduces singularities.

4.1.1.2 Visual data modalities

As already discussed in Section 2.3 to ground the robots actions in the physical world, they are dependent on sensory input. Here, in the majority of methods visual observations are used. In the original work (Chi et al., 2023), combining visual robotic manipulation with DMs for trajectory planning, the DM is conditioned on RGB-image observations. Many methods, e.g., (Si et al., 2024; Pearce et al., 2022; Li X. et al., 2024), adopt using RGB inputs, also developing more intricate encoding schemes (Qi et al., 2025).

However, 2D visual scene representations may not provide sufficient geometrical information for intricate robotic tasks, especially in scenes containing occlusions. Thus, multiple later methods used 3D scene representations instead. Here, DMs are either directly conditioned on the point cloud (Li et al., 2025; Liu et al., 2023c; Wang C. et al., 2024) or point cloud feature embeddings (Ze et al., 2024; Xian et al., 2023; Ke et al., 2024), from singleview (Ze et al., 2024; Li et al., 2025; Wang C. et al., 2024), or multiview camera setups (Ke et al., 2024; Xian et al., 2023). While multiview camera setups provide more complete scene information, they also require a more involved setup and more hardware resources.

These models outperform methods relying solely on 2D visual information, on more complex tasks, also demonstrating robustness to adversarial lighting conditions.

4.1.1.3 Trajectory planning as image generation

Another category formulates trajectory generation directly in image space, leveraging the exceptional generative abilities of DMs in image generation. Here (Ko et al., 2024; Zhou S. et al., 2024; Du et al., 2023), given a single image observation, a sequence of images, or a video, sometimes in combination with a language-task-instruction, the diffusion process is conditioned to predict a sequence of images, depicting the change in robot and object position. This comes with the benefit of internet-wide video training data, which facilitates extensive training, leading to good generalization capabilities. Especially in combination with methods (Bharadhwaj et al., 2024b) agnostic to the robot embodiment, this highly increases the amount of available training data. Moreover, in robotic manipulation, the model usually has to parse visual observations. Predicting actions in image space circumvents the need for mapping from the image space to a usually much lower-dimensional action space, reducing the required amount of training data (Vosylius et al., 2024). However, predicting high-dimensional images may also prevent the model from successfully learning important details of trajectories, as the DM is not guided to pay more attention to certain regions of the image, even though usually only a low fraction of pixels contain task-relevant information. Additionally, methods generating complete images must ensure temporal consistency and physical plausibility. Hence, extensive training resources are required. As an example (Zhou S. et al., 2024), uses 100 V100 GPUs and 70k demonstrations for training. While still operating in image space, some methods do not generate whole image sequences, but instead perform point-tracking (Bharadhwaj et al., 2024b) or diffuse imprecise action-effects on the end-effector position directly in image space (Vosylius et al., 2024). This mitigates the problem of generating physically implausible scenes. However, point-tracking still requires extensive amounts of data. Bharadhwaj et al. (2024b), e.g., uses 0.4 million video clips for training.

4.1.1.4 Long-horizon and multi-task learning

Due to their ability to robustly model multi-model distributions and relatively good generalization capabilities, DMs are well suited to handle long-horizon and multi-skill tasks, where usually long-range dependencies and multiple valid solutions exist, especially for high-level task instructions (Mendez-Mendez et al., 2023; Liang et al., 2024). Often, long-horizon tasks are modeled using hierarchical structures and skill learning. Usually, a single skill-conditioned DM or several DMs are learned for the individual skills, while the higher-level skill planning does not use a DM (Mishra et al., 2023; Kim W. K. et al., 2024; Xu et al., 2023; Liang et al., 2024; Li et al., 2023). The exact architecture for the higher-level skill planning varies across methods, being, for example, a variational autoencoder (Kim W. K. et al., 2024) or a regression model (Mishra et al., 2023). Instead of having a separate skill planner that samples one skill, Wang L. et al. (2024) develops a sampling scheme that can sample from a combination of DMs trained for different tasks and in different settings.

To forego the skill-enumeration, which brings with it the limitation of a predefined finite number of skills, some works employ a coarse-to-fine hierarchical framework, where higher-level policies are used to predict goal states for lower-level policies (Zhang E. et al., 2024; Ma X. et al., 2024; Xian et al., 2023; Ha et al., 2023; Huang Z. et al., 2024; Du et al., 2023).

The ability of DMs to stably process high-dimensional input spaces enables the integration of multi-modal inputs, which is especially important in multi-skill tasks, to develop versatile and generalizable agents via arbitrary skill-chaining. Methodologies use videos (Xu et al., 2023), images, and natural language task instructions (Liang et al., 2024; Wang L. et al., 2024; Zhou S. et al., 2024; Reuss et al., 2024b), or even more diverse modalities, such as tactile information and point clouds (Wang L. et al., 2024), to prompt skills.

Although these methods are designed to enhance generalizability, achieving adaptability in highly dynamic environments and unfamiliar scenarios may require the integration of continuous and lifelong learning. This is a widely unexplored field in the context of DMs, with only very few works (Huang J. et al., 2024; Di Palo et al., 2024) exploring this topic. Moreover, these methods are still limited in their applications. Di Palo et al. (2024) are utilizing a lifelong buffer to accelerate the training of new policies for new tasks. In contrast, Mendez-Mendez et al. (2023) continually updates its policy. However, they only conduct training and experiments in simulation. Additionally, their method requires precise feature descriptions of all involved objects and is limited to predefined abstract skills. Moreover, for the continual update, all past data is replayed, which is

not only computationally inefficient but also does not prevent catastrophic forgetting.

4.1.1.5 Multi-task learning with vision language action models

Another approach to enhance generalizability in multi-task settings is the incorporation of pretrained VLAs. As a specialized class of multimodal language model (MLLM), VLAs combine the perceptual and semantic representation power of the vision language foundation model and the motor execution capabilities of the action generation model, thereby forming a cohesive end-to-end decision-making framework. Being pretrained on internet-scale data, VLAs exhibit great generalization capabilities across diverse and unseen scenarios, thereby enabling robots to execute complex tasks with remarkable adaptability (Firoozi et al., 2025).

A predominant line of approaches among VLAs employs next-token prediction for auto-regressive action token generation, representing a foundational approach to end-to-end VLA modeling, e.g., (Brohan et al., 2023b; Brohan et al., 2023a; Kim M. J. et al., 2024). However, this approach is hindered by significant limitations, most notably the slow inference speeds inherent to auto-regressive methods (Brohan et al., 2023a; Wen et al., 2025; Pertsch et al., 2025). This poses a critical bottleneck for real-time robotic systems, where low-latency decision-making is essential. Furthermore, the discretizations of motion tokens, which reformulates action generation as a classification task, introduces quantization errors that lead to a decrease in control precision, thus reducing the overall performance and reliability (Zhang et al., 2024g; Pearce et al., 2022; Zhang S. et al., 2024).

To address these limitations one line of research within VLAs focuses on predicting future states and synthesizing executable actions by leveraging inverse kinematics principles derived from these predictions, e.g., (Cheang et al., 2024; Zhen et al., 2024; Zhang et al., 2024c). While this approach addresses some of the limitations associated with token discretization, multimodal states often correspond to multiple valid actions, and the attempt to model these states through techniques such as arithmetic averaging can result in infeasible or suboptimal action outputs.

Thus, showing strong capabilities and stability in modeling multi-modal distributions, DMs have emerged as a promising solution. Leveraging their strong generalization capabilities, a VLA is used to predict coarse action, while a DM-based policy refines the action, to increase precision and adaptability to different robot embodiments, e.g. (Pan C. et al., 2024; Shentu et al., 2024; Team et al., 2024). For instance, TinyVLA (Wen et al., 2025) incorporates a diffusion-based head module on top of a pretrained VLA to directly generate robotic actions. More specifically, DP (Chi et al., 2023) is connected to the multimodal model backbone via two linear projections and a LayerNorm. The multimodal model backbone jointly encodes the current observations and language instruction, generating a multimodal embedding that conditions and guides the denoising process. Furthermore, in order to better fill the gap between logical reasoning and actionable robot policies, a reasoning injection module is proposed, which reuses reasoning outputs (Wen et al., 2024). Similarly, conditional diffusion decoders have been leveraged to represent continuous multimodal action distributions, enabling the generation of diverse and contextually appropriate action sequences (Team et al., 2024; Liu et al., 2024; Li Q. et al., 2024).

Addressing the disadvantage of long inference times with DMs, in some recent works instead, flow matching is used to generate actions from observations preprocessed by VLMs to solve flexible and dynamic tasks, offering a robust alternative to traditional diffusion mechanisms (Black et al., 2024a; Zhang and Gienger, 2025). While Black et al. (2024a) takes a skill-based approach, where the vision-language model is used to decide on actions, Zhang and Gienger (2025) uses a vision-language model to generate waypoints. In both approaches, flow matching is used as the expert policy, generating precise trajectories.

VLAs offer access to models trained on huge amounts of data and with strong computational power, leading to strong generalization capabilities. To mitigate some of their shortcomings, such as imprecise actions, specialized policies can be used for refinement. To not restrict the generalizability of the VLA, DMs offer a great possibility, as they can capture complex multi-model distributions and process high-dimensional visual inputs. However, both VLAs and DMs have a relatively slow inference speed. Thus, especially in this combination with VLAs, increasing the sampling efficiency of DMs is important. One example was provided in the previous paragraph. But the topic of higher sampling speed with DMs is also discussed in more detail in Section 2.2.1.

4.1.1.6 Constrained planning

Another line of methods focuses on constrained trajectory learning. A typical goal is obstacle avoidance, object-centric, or goal-oriented trajectory planning, but other constraints can also be included. If the constraints are known prior to training, they can be integrated into the loss function. However, if the goal is to adhere to various and possibly changing constraints during inference, another approach has to be taken. For less complex constraints, such as specific initial or goal states (Janner et al., 2022), introduces a conditioning, where, after each denoising time step (Equation 7), the particular state from the trajectory is replaced by the state from the constraint. However, this can lead the trajectory into regions of low likelihood, hence decreasing stability and potentially causing mode collapse. Moreover, this method is not applicable to more complex constraints.

One approach, also addressed by Janner et al. (2022), is classifier guidance (Dhariwal and Nichol, 2021). Here, a separate model is trained to score the trajectory at each denoising step and steer it toward regions that satisfy the constraint. This is integrated into the denoising process by adding the gradient of the predicted score. It should be noted that for sequential data, such as trajectories, classifier guidance can also bias the sampling towards regions of low likelihood (Pearce et al., 2022). Thus, the weight of the guidance factor must be carefully chosen. Moreover, during the start of the denoising process the guidance model must predict the score on a highly uninformative output (close to Gaussian noise) and should have a lower impact. Therefore, it is important to inform the classifier of the denoising time step, train it also on noisy samples, or adjust the weight with which the guidance factor is integrated into the reverse process. Classifier guidance is applied in several methodologies (Mishra et al., 2023; Liang et al., 2023; Janner et al., 2022; Carvalho et al., 2023). However, it requires the additional training of a separate model. Furthermore,

computing the gradient of the classifier at each sampling step adds additional computational cost. Thus, classifier-free guidance (Ho et al., 2021; Saha et al., 2024; Li et al., 2025; Power et al., 2023; Reuss et al., 2024a; Reuss et al., 2023) has been introduced, where a conditional and an unconditional DM per constraint are trained in parallel. During sampling, a weighted mixture of both DMs is used, allowing for arbitrary combinations of constraints, also not seen together during training. However, it does not generalize to entirely new constraints, as this would necessitate the training of new conditional DMs.

As both classifier and classifier-free guidance only steer the training process, they do not guarantee constraint satisfaction. To guarantee constraint satisfaction in delicate environments, such as surgery (Scheikl et al., 2024), incorporate movement primitives with DMs to ensure the quality of the trajectory. Recent advances in diffusion models also delve into constraint satisfaction (Römer et al., 2024), integrating constraint tightening into the reverse diffusion process. While this outperforms previous methods (Power et al., 2023; Janner et al., 2022; Carvalho et al., 2024) in regards to constraint satisfaction, also in multi-constraint settings and constraints not seen during training, the evaluation is done only in simulation on a single experiment setup. Thus, constraint satisfaction with DMs remains an interesting research direction to further explore.

Few methods also perform affordance-based optimization for trajectory planning (Liu et al., 2023c). However, most work in affordance-based manipulation concentrates on grasp learning, which is discussed in more detail in Section 4.2.

4.1.2 Offline reinforcement learning

To apply diffusion policies in the context of RL the reward term has to be integrated. Diffuser (Janner et al., 2022), one early work adapting diffusion to RL, uses classifier-based guidance, which is based on classifier guidance described in Section 4.1.1.6. Let $\tau = \{(s_0, a_0), \dots, (s_T, a_T)\}$ be a trajectory with one state-action pair per timestep in a planning horizon $\{0, \dots, T\}$. To incorporate the reward term during sampling, a regression model $R_\phi(\tau_k)$ is trained to predict the return, i.e., the cumulative future reward, over the trajectory τ_k at each denoising time step $k \in \{0, \dots, K\}$. This is incorporated into the sampling process by adding the guidance term at each iteration of the reverse diffusion process (Janner et al., 2022):

$$p\left(\tau_{k-1} \mid \tau_k, \mathcal{O}_{1:T}\right) \approx \mathcal{N}\left(\tau_{k-1}; \mu + \Sigma \nabla R_\phi\left(\mu\right), \Sigma\right). \tag{8}$$

Moreover, to ensure that the current state observation s_0 is not changed by the reverse diffusion on the trajectory, $\tau_{s_0}^{k-1}$ is set to the current state observation after each reverse diffusion iteration. In the same way, goal-conditioning or other constraints, which can be accomplished by replacing states from the trajectory with states from the constraint, can be integrated into the method. This, is done in several methodologies (Janner et al., 2022; Liang et al., 2023). However, it has to be done with care, as it can lead to trajectories in regions of low likelihood which may cause instability and mode-collapse (Janner et al., 2022; Song Y. et al., 2021). After the reverse process is completed and τ^0 has been predicted, the first action a_0 of the plan is executed. Then, the planning horizon is shifted one step forward, and the next action is sampled.

In Diffuser (Janner et al., 2022) and Diffuser-based methods (Suh et al., 2023; Liang et al., 2023), the DM is trained independently of the reward signal, similar to methods in imitation learning with DM. Not leveraging the reward signal for training the policy can lead to misalignment of the learned trajectories with optimal trajectories and thus suboptimal behavior of the policy. In contrast, leveraging the reward signal already during training of the policy, can steer the training process, consequently increasing both quality of the trained policy and sample efficiency.

To mitigate these shortcomings, one approach, Decision Diffuser (Ajay et al., 2023), directly conditions the DM on the return of the trajectory using classifier-free guidance. This method outperforms Diffuser on a variety of tasks, such a block-stacking task. However, both methods have not been evaluated on real-world tasks. Directly conditioning on the return, limits generalization capabilities. Different to Q-learning, where the value function is approximated, which generalizes across all future trajectories, here only the return of the current trajectory is considered. Sharing some similarity to on-policy methods, this limits generalization as the policy learns to follow trajectories from the demonstrations with high return values. Thus, this can also be interpreted as guided imitation learning.

A more common method (Wang et al., 2023a) integrates offline Q-learning with DMs. The loss function from Equation 5 is a behavior cloning loss, as the goal is to minimize error with respect to samples taken via the behavior policy. Wang et al. (2023a) suggests including a critic in the training procedure, which they call Diffusion Q-learning (Diffusion-QL). In Diffusion-QL a Q-function is trained, by minimizing the Bellman-Operator using the double Q-learning trick. The actions for updating the Q-function are sampled from the DM. In turn a policy improvement step $\mathcal{L}_c = -\mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}, \boldsymbol{a}^0\sim\pi_\theta}\left[Q_\phi(\boldsymbol{s}, \boldsymbol{a}^0)\right]$ is included in the loss for updating the DM (Wang et al., 2023a):

$$\pi = \arg\min_{\pi_\theta} \mathcal{L}_{RL} = \arg\min_{\pi_\theta} \mathcal{L} + \alpha \mathcal{L}_c, \tag{9}$$

where $\mathcal{L}$ is the diffusion loss from Equation 5 and the parameter α regulates the influence of the critic. Several methods (Ada et al., 2024; Kim S. et al., 2024; Venkatraman et al., 2023; Kang et al., 2023), build on Diffusion Q-learning. To increase the generalizability to out-of-distribution data, a common problem in offline RL (Levine et al., 2020), Ada et al. (2024), include a state-reconstruction loss, into the training of the DM. An overview of the architectures of methods combining diffusion and reinforcement learning is provided in Table 3.

One characteristic of methodologies combining RL with DMs is that they are offline methods, with both the policy, i.e., the DM, and the return prediction model/critic being trained offline. This introduces the usual advantages and disadvantages of offline RL (Levine et al., 2020). The model relies on high-quality existing data, consisting of state-action-reward transitions, and is unable to react to distribution shifts. If not tuned well, this may also lead to overfitting. On the other hand, it has increased sample efficiency and does not require real-time data collections and training, which decreases computational cost and can increase training stability. Compared to imitation learning (Levine et al., 2020; Pfrommer et al., 2024; Ho and Ermon, 2016), offline RL requires data labeled with rewards, the training of a reward function, and is more prone to overfitting to suboptimal behavior. However, confronted with data containing diverse and suboptimal behavior, offline RL has the potential of better generalization compared to

imitation learning, as it is well suited to model the entire state-action space. Thus, combining RL with DMs has the potential of modeling highly multi-modal distributions over the whole state-action space, strongly increasing generalizability (Liang et al., 2023; Ren et al., 2024). In contrast, if high-quality expert demonstrations are available, imitation learning might lead to better performance and computational efficiency. To overcome some of the shortcoming of imitation learning, such as the covariate shift problem (Ross and Bagnell, 2010), which make it difficult to handle out of distribution situations, some strategies are devised to finetune behavior cloning policies using RL (Ren et al., 2024; Huang T. et al., 2025).

Skill-composition is a common method, to handle long-horizon tasks. To leverage the abilities of RL to learn from suboptimal behaviors multiple methodologies (Ajay et al., 2023; Kim W. K. et al., 2024; Venkatraman et al., 2023; Kim S. et al., 2024) combine skill-learning and RL with DMs.

Only little research (Ding and Jin, 2023; Ajay et al., 2023) in online and offline-to-online RL with DMs has been conducted, leaving a wide field open for research. Moreover, in the context of skill-learning (Ajay et al., 2023), the DMs, used for the lower-level policies, are trained offline and remain frozen, while the higher-level policy are trained using online RL.

It should be noted that, apart from Ren et al. (2024); Huang T. et al. (2025), none of the aforementioned methods process visual observations and instead rely on ground-truth environment information, which is only easily available in simulation. Moreover, while all methods have also been tested on robotic manipulation tasks, only a few (Ren et al., 2024; Huang T. et al., 2025) have been deliberately engineered for these specific applications. Expanding the scope to encompass all methodologies devised for robotics at large, there is a more substantial body of work that integrates diffusion policies with RL.

4.2 Robotic grasp generation

Grasp learning, as one of the crucial skills for robotic manipulation, has been studied over decades (Newbury et al., 2023). Starting from hand-crafted feature engineering to statistical approaches (Bohg et al., 2013), accompanied by the recent progress in deep neural networks that are powered by massive data collection either from real-world (Fang et al., 2020) or simulated environments (Gilles et al., 2023; Gilles et al., 2025; Shi et al., 2024). The current trend in grasp learning incorporates semantic-level object detection, leveraging open-vocabulary foundation models (Radford et al., 2021; Liu et al., 2025), and focuses on object-centric or affordance-based grasp detection in the wild (Qian et al., 2024; Shi et al., 2025). To this end, DMs, known for their ability to model complex distributions, allow for the creation of diverse and realistic grasp scenarios by simulating possible interactions with objects in a variety of contexts (Rombach et al., 2022b). Furthermore, these models contribute to direct grasp generation by optimizing the generation of feasible and efficient grasps (Urain et al., 2023), particularly in environments where real-time decision-making and adaptability are critical.

Grasp generation with DMs can be categorized into several key approaches: From methodological perspective, one category focuses on explicit diffusion on 6-DoF grasp poses that lie on the $\mathbf{SE}(3)$ group, directly modeling spatial transformations to generate feasible grasps (Urain et al., 2023; Song et al., 2024; Wu et al., 2024b; Weng et al., 2024; Singh et al., 2024; Lim et al., 2024). Another line of approaches involves implicit grasp diffusion within latent space, enhancing adaptability and versatility (Barad et al., 2024). A recent trend focuses on language-guided diffusion for task-oriented grasp generation, where natural language inputs shape the generation process (Nguyen N. et al., 2024; Vuong et al., 2024; Nguyen T. et al., 2024; Chang and Sun, 2024). Other approaches emphasize affordance-driven diffusion, targeting specific functional goals, such as object pose diffusion for rearrangement (Liu et al., 2023b; Zhao et al., 2025), affordance-guided object reorientation (Mishra and Chen, 2024), imitation learning (Wu et al., 2024a; Ma C. et al., 2024) or multi-embodiment grasping (Freiberg et al., 2025). Apart from these categories, hand-object interaction (HOI) specifically prioritizes the synthesis of realistic, functional interactions by modeling the hand's adaptive responses to various object shapes and affordances with dexterity (Ye et al., 2024; Wang Y.-K. et al., 2024; Zhang et al., 2024d; Cao et al., 2024; Li P. et al., 2024; Zhang et al., 2025; Lu et al., 2025; Zhang et al., 2024b). In addition to the diffusion on grasp generation or trajectory planning, DM as sim-to-real generator (Li Y. et al., 2024) or foundational feature extractor (Tsagkas et al., 2024) such as stable diffusion (Rombach et al., 2022a) may provide semantic information to enhance downstream grasp generation tasks. Table 4 summarizes the aforementioned categories. Notably, we include the applications of diffusion in HOI, imitation learning for pre-grasp, and tasks related to image generation in the graph, which will not be further discussed in the rest of this survey due to their relevance to the field of computer vision. While readers are still encouraged to refer to the relevant literature according to our illustration (Table 4: HOI Synthesis). More details on the architectures of the individual methods in grasp learning are provided in Table 5.

4.2.1 Diffusion as $\mathbf{SE}(3)$ grasp pose generation

Since the standard diffusion process is primarily formulated in Euclidean space, directly extending it to $\mathbf{SE}(3)$ poses, represented by: $\mathbf{H} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ is inherently challenging due to potential numerical instability (to satisfy $\mathbf{H}\mathbf{H}^{-1} = \mathbf{I}^{4\times4}$), since typical Langevin dynamics cannot be applied for non-Euclidean manifolds such as the $\mathbf{SE}(3)$ Lie group. Here, $\mathbf{R} \in \mathbf{SO}(3)$ represents the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ the translation vector. Applying diffusion to $\mathbf{SE}(3)$ poses requires accounting for the manifold's non-Euclidean nature, where standard Gaussian noise, as used in vanilla diffusion, fails to retain stability over rotations and translations.

To tackle this, $\mathbf{SE}(3)$-Diff (Urain et al., 2023) introduced a smooth cost function to learn the grasp quality via the energy-based model (EBM), where the score matching for EBM is applied on the Lie group to bridge the gap between diffusion processes on the vector space $\mathbb{R}^6$ and the $\mathbf{SE}(3)$. In contrast, Song et al. (2024) condition the 6-Dof grasp poses on the grasp locations $\mathbf{t}$ and corresponding volumetric features for grasp generation in clutter following GIGA framework (Jiang et al., 2021), without explicit consideration on the $\mathbf{SE}(3)$ constraint. Moreover, one advantage of the EBM model in $\mathbf{SE}(3)$-Diff is the direct grasp quality evaluation and integration into the entire grasp motion planning and optimization. However, training EBM-based models demands extensive sampling

TABLE 4 Taxonomy of grasp generation approaches with diffusion models.

Perspective	Category	Subcategory	References
Methodological	Diffusion on **SE**(3) grasp poses	Parallel jaw grasp	Urain et al. (2023), Song et al. (2024), Singh et al. (2024), Lim et al. (2024), Carvalho et al. (2024), Ryu et al. (2024), Freiberg et al. (2025), Huang et al. (2025a)
		Dextrous grasp	Wu et al. (2024b), Weng et al. (2024), Wang et al. (2024c), Freiberg et al. (2025), Zhong and Allen-Blanchette (2025), Zhang et al. (2024h), Wu et al. (2023)
	Diffusion in latent space	-	Barad et al. (2024)
	Diffusion as feature encoders and image generators	-	Li et al. (2024d), Tsagkas et al. (2024)
Functional	Affordance-driven diffusion	Language-guided grasp diffusion	Nguyen et al. (2024a), Vuong et al. (2024), Nguyen et al. (2024b), Chang and Sun (2024), Zhang et al. (2025)
		Pre-grasp manipulation via imitation learning	Wu et al. (2024a), Ma et al. (2024a)
	HOI synthesis	-	Ye et al. (2024), Wang et al. (2024c), Zhang et al. (2024d), Cao et al. (2024), Li et al. (2024a), Zhang et al. (2025), Lu et al. (2025)
	Object pose diffusion for reorientation and rearrangement	-	Liu et al. (2023b), Simeonov et al. (2023), Mishra and Chen (2024), Zhao et al. (2025)

and poses significant challenges for generalization. We noticed that flow matching (Lipman et al., 2023) is employed in recent studies, such as EquiGraspFlow (Lim et al., 2024) and Grasp Diffusion Network (Carvalho et al., 2024), which use continuous normalizing flows (CNFs) as ODE solvers to learn angular (**SO**(3)) and linear (**R**(3)) velocities for denoising. This preserves the **SE**(3)-equivariance conditioned on the input point cloud given the time schedule. In contrast to **SE**(3)-Diff, which relies on additional supervision in the form of signed distance functions, they achieve competitive performance without requiring this auxiliary module, leading to more efficient training. In general, although CNF-based approaches exhibit promising performance on grasp generation for a single object, more studies on generalizability to highly occluded environments (Freiberg et al., 2025) and uncertainty quantification (Shi et al., 2024) are expected in future work.

In contrast to explicit pose diffusion, latent DMs for grasp generation (GraspLDM (Barad et al., 2024)) explore latent space diffusion with VAEs, which does not explicitly account for the **SE**(3) constraint. They follow VAE-based 6-Dof Graspnet (Mousavian et al., 2019) to model the distribution of grasp latent features by a denoising diffusion process, which is conditioned on the point cloud and task latent for the grasp generation. This implicit modeling may potentially limit the model's ability to generate physically plausible and geometrically consistent grasp poses.

Furthermore, the **SE**(3) bi-equivariance property is critical for efficient grasp generation (Huang et al., 2023), as it requires that any transformation applied to the input space correspondingly transforms the output space in a consistent manner. Specifically, this property implies that the generated poses from a **SE**(3)-invariant distribution should maintain the same spatial and geometric relationships under transformations over the time schedule, ensuring that the learned grasp distribution remains invariant across various orientations and positions. For instance, Ryu et al. (2023) consider bi-equivariance in Lie group representation to construct the equivariance descriptor field (EDF) (Ryu et al., 2023), taking the transformations of both observation (target) space and initial end-effector frame into account. This principally improves the sample efficiency on pick-and-place tasks via Imitation learning. Upon this, they extend the EDF to bi-equivariant score matching (Ryu et al., 2024) to be applied in the context of diffusion, which consists of both translational and rotational fields on **se**(3) Lie algebra. Moreover, Freiberg et al. (2025) adapts the approach from Ryu et al. (2024) to generalize to multi-embodiment grasping through an equivariant encoder that captures gripper embeddings. In terms of the theoretical background to equivariant robot learning, we identify a recent survey (Seo et al., 2025) as a recommendation for interested readers.

4.3 Visual data augmentation

One line of methodologies focuses on employing mostly pretrained DMs for data augmentation in vision-based manipulation tasks. Here, the strong image generation and processing capabilities of diffusion generative models are utilized

TABLE 5 Technical details of grasp diffusion methodologies on SE(3) grasp synthesis. The references for the encoders are provided in Supplementary Appendix Table 1. The references for the benchmarks are listed in Supplementary Appendix Table 2. In the following, the abbreviations used are explained: SDF: Signed Distance Function, TSDF: Truncated SDF, PCs: Point Clouds, FiLM: Convolutional Neural Network with Feature-wise Linear Modulation (Perez et al., 2018), DiTs: Diffusion Transformers, Eq.: Equivariant, VN: Vector Neuron.

Reference	Input	Encoder	Diffuser	Benchmark
Urain et al. (2023)	SDF	Shape encoder	FiLM	Acronym
Barad et al. (2024)	PCs	PointNet++	FiLM	Acronym
Song et al. (2024)	TSDF	OccNet	FiLM	VGN
Singh et al. (2024)	PCs	OccNet	FiLM	DA^2
Lim et al. (2024)	PCs	VN-DGCNN	FiLM	Acronym
Freiberg et al. (2025)	PCs + Gripper PCs	Eq. U-Net	Eq. FiLM	Self generated
Carvalho et al. (2024)	PCs	PointNet++	DiT	Acronym
Huang et al. (2025a)	PCs + Guidance	VN-PointNet	DiTs	OakInk
Weng et al. (2024)	PCs + Gripper PCs	BPS	DiTs	DexGraspNet
Zhong and Allen-Blanchette (2025)	PCs + Gripper PCs	Eq. Models	Eq. DiTs	MultiDex
Zhang et al. (2024h)	PCs	PointNet++	DiTs	MultiDex

to augment data sets and scenes. The main goals of the visual data augmentation are scaling up data sets, scene reconstruction, and scene rearrangement.

4.3.1 Scaling data and scene augmentation

A challenge associated with data-driven approaches in robotics relates to substantial data requirements, which are time-consuming to acquire, particularly for real-world data. In the domain of imitation learning, it is essential to accumulate an adequate number of expert demonstrations that accurately represent the task at hand. While, by now, many methods, e.g., (Reuss et al., 2024a; Ze et al., 2024; Ryu et al., 2024) only require a low number of five to fifty demonstrations, there are also methods, e.g., (Chen L. et al., 2023; Saha et al., 2024) relying on more extensive data sets. Especially offline RL methods, e.g. (Carvalho et al., 2023; Ajay et al., 2023) usually require extensive amounts of data to accurately predict actions over the complete state-action space, also from suboptimal behavior. Moreover, increasing the variability in training data also has the potential to increase the generalizability of the learned policies. Thus, to automatically increase the variety and size of datasets, without additional costs on researchers and staff, or other more engineering-heavy autonomous data collection pipelines (Yu et al., 2023), many methodologies, e.g., (Chen Z. et al., 2023; Mandi et al., 2022), use DMs for data augmentation. In comparison to other strategies, such as domain randomization (Tremblay et al., 2018; Tobin et al., 2017), data augmentation with DMs directly augments the real-world data, making the data grounded in the physical world. In contrast, domain randomization requires complex tuning for each task, to ensure physical plausibility of the randomized scenes, and to enable sim-to-real transfer (Chen Z. et al., 2023).

Given a set of real-world data, DM-based augmentation methods perform semantically meaningful augmentations via inpainting, such as changing object colors and textures (Zhang X. et al., 2024), or even replacing whole objects, as well as corresponding language task descriptions (Chen Z. et al., 2023; Yu et al., 2023; Mandi et al., 2022). This enables both the augmentation of objects, which are part of the manipulation process, and backgrounds. The former increases the generalizability to different tasks and objects, while the latter increases robustness to scene information, which should not influence the policy. Some (Zhang X. et al., 2024) also augment object positions and the corresponding trajectories to generate off-distribution demonstrations for DAgger, thus addressing the covariate shift problem in imitation learning. Others (Chen L. Y. et al., 2024) augment camera view, robot embodiments, or even (Katara et al., 2024) generate whole simulation scenes from given URDF files, prompted by a Large Language Model (LLM). Targeted towards offline RL methods, Di Palo et al. (2024) combines data augmentation with a form of hindsight-experience replay (Andrychowicz et al., 2017) to adapt the visual observations to the language-task instruction. This increases the number of successful executions in the replay buffer, which potentially increases the data efficiency. The method is used to learn policies for new tasks, on previously collected data, to align the data with the new task instructions.

From a methodological perspective the methods mostly employ frozen web-scale pretrained language (Yu et al., 2023), and vision-language models, for object segmentation (Yu et al., 2023), or text-to-image synthesis (Stable Diffusion) (Rombach et al., 2022a; Mandi et al., 2022), or finetune (Zhang X. et al., 2024; Di Palo et al., 2024) pretrained internet-scale vision-language models. Apart from Zhang X. et al. (2024) the methods,

do not augment actions, but only observations. Thus, the methodologies must ensure augmentations, for which the demonstrated actions do not change, which highly limits the types of augmentations. Moreover, large-scale data scaling via scene augmentation also requires additional computational cost. While this might not be a severe limitation, if it is applied once before the training, it may highly increase training time for online-RL methods.

4.3.2 Sensor data reconstruction

A challenge in vision-based robotic manipulation pertains to the incomplete sensor data. Especially single-view camera setups lead to incomplete object point clouds or images, making accurate grasp and trajectory prediction challenging. This is exacerbated by more complex task settings, with occlusion, as well as inaccurate sensor data.

Multiple methods (Kasahara et al., 2024; Ikeda et al., 2024) reconstruct camera viewpoints with DMs. Given an RGBD image and camera intrinsics Kasahara et al. (2024) generates new object views without requiring CAD models of the objects. For this, the existing points are projected to the new viewpoint. The scene is segmented using the vision foundation model SAM (Kirillov et al., 2023), to create object masks. On these masks missing data points are inpainted using the pretrained diffusion model for image generation Dall·E (Kapelyukh et al., 2023). As Dall·E does not ensure spatial consistency, consistency filtering is applied across viewpoints. Moreover, Dall·E, only processes 2D images. Thus, to also complete the missing depth information, a model is trained to predict the missing depth information from the projected depth map and the reconstructed image. In this method the viewpoints are sampled on evenly spaced directions along a viewing sphere. However, generating the point clouds for many viewpoints is computationally expensive, and might not be necessary for successful task completion. Thus, view-planning is applied to generate a minimal set of views Pan S. et al. (2024), Pan et al. (2025). use a DM to generate geometric priors from a 2D image, enabling a view-planner to sample a minimum set of viewpoints that minimize movement cost. The views are then used to train a Neural Radiance Field (NeRF) (Mildenhall et al., 2020) to reconstruct 3D scenes from 2D images.

In the field of robotic manipulation, not many methods consider scene reconstruction. A possible reason for this is its relatively high computational cost. However, expanding to the areas of robotics and computer vision, more methodologies in the field of scene reconstruction exist. In robotic manipulation instead more methods focus on making policies more robust to incomplete or noisy sensor information, e.g., (Ze et al., 2024; Ke et al., 2024). However, the limited number of occlusion in the experimental setups indicate that strong occlusion are still a major challenge. Moreover, scene reconstruction is unable to react to completely occluded objects.

4.3.3 Object rearrangement

The ability of DMs for text-to-image synthesis offers the possibility to generate plans from high-level task descriptions. In particular, given an initial visual observation, one group of methods uses such models to generate target-arrangement of objects in the scene, specified by a language-prompt (Liu et al., 2023b; Kapelyukh et al., 2023; Xu et al., 2024; Zeng et al., 2024; Kapelyukh et al., 2024). Examples of applications could be setting up a dinner table or clearing up a kitchen counter. While the earlier methodologies (Kapelyukh et al., 2023; Liu et al., 2023b) use the pretrained VLM Dall·E (Black et al., 2024b) to generate rearrangements in a zero-shot manner, this has the disadvantage of possibly introducing scene inconsistencies and incompatibilities, due to the lack of geometric understanding and object permanence. Thus, the later methods (Xu et al., 2024; Kapelyukh et al., 2024) use combinations of pretrained LLMs and VLMs like CLIP (Meila and Zhang, 2021), together with other non-diffusion visual processing methods like NeRF (Mildenhall et al., 2020) and SAM (Kirillov et al., 2023), and custom DMs. The described methodologies are similar to the methods for object pose diffusion (Mishra and Chen, 2024; Simeonov et al., 2023; Zhao et al., 2025) mentioned in Section 4.2. The main difference is that the methods here focus on the rearrangement of multiple objects specified by a sparse language input, not exhaustively describing the geometric layout of the target arrangement. Different to the methods from Section 4.2, the integration with grasp or motion planning to achieve the target arrangement is not the focus. However, nonetheless for all of the above listed methodologies for object rearrangement their effectiveness is also demonstrated in real-robot experiments.

5 Experiments and benchmarks

In this section, we focus on the evaluation of the various DMs for robotic manipulation. Details on the employed benchmarks and baselines are listed in the separate tables for imitation learning (Table 6), reinforcement learning (Table 7) in the Appendix, and grasp learning (Table 5). Separately, the references for all applied benchmarks are listed in Supplementary Appendix Table 2.

Various benchmarks are used to evaluate the methods. Common benchmarks are CALVIN (Mees et al., 2022), RLBench (James et al., 2020), RelayKitchen (Gupta et al., 2020), and Meta-World (Yu et al., 2020). Primarily in RL, the benchmark D4RL Kitchen (Fu et al., 2020) is used. One method (Ren et al., 2024) uses FurnitureBench (Heo et al., 0) for real-world manipulation tasks. Adroit (Rajeswaran et al., 2017) is a common benchmark for dexterous manipulation, LIBERO (Liu B. et al., 2023) for lifelong learning, and LapGym (Maria Scheikl et al., 2023) for medical tasks.

Many methods are only being evaluated against baselines, which are not based on DMs themselves. However, there are some common DM-based baselines. For methods operating in **SE**(3)-space (Chen K. et al., 2024; Song et al., 2024; Ryu et al., 2024), **SE**(3)-Diffusion Policy (Urain et al., 2023), probably the first paper using DMs for grasp generation, is commonly used as baseline. For RL-based methods, the RL-based Diffuser (Janner et al., 2022), Diffusion-QL (Wang et al., 2023a), and Decision Diffuser (Ajay et al., 2023) are commonly used as baselines. It should be noted that in the original paper, Decision Diffuser (Ajay et al., 2023) is evaluated against Diffuser (Janner et al., 2022) and outperforms it on almost all tasks, particularly on the manipulation tasks, block stacking, and rearrangement. However, neither of these methods is evaluated on real-world tasks. Another common baseline is DP

TABLE 6 Benchmarks of trajectory diffusion using reinforcement learning.

Reference	Diffusion Baseline	Simulation		Real World	
		Benchmark	#Demos	Real	#Demos
Diffuser (Janner et al., 2022)	✗	KUKA (custom)	10 k	✗	-
Decision Diffuser (Ajay et al., 2023)	✗	D4RLKitchen, KUKA	/, 10 k	✗	-
Diffusion-QL (Wang et al., 2023b)	✗	D4RLKitchen	10000 trans*1	✗	-
Wang et al. (2023b)	Diffuser	custom	8 k	✗	-
HDMI (Li et al., 2023)	✗	✓	-	✗	-
Ding and Jin (2023)	Diffusion-QL	D4RLKitchen, Adroit	/	✗	-
Mishra et al. (2023)	Decision Diffuser	STAP	/	✓	/
Kang et al. (2023)	Diffusion-QL	Adroit, D4RL Kitchen	/	✗	-
Brehmer et al. (2023)	Diffuser	KUKA	/	✗	-
Suh et al. (2023)	Diffuser	✓	-	✓	100
Ha et al. (2023)	Diffusion-QL	✗	-	✓	50
Kim et al. (2024b)	Diffuser, Decision Diffuser, HDMI	Fetch env	/	✗	-
Liang et al. (2023)	Diffuser, Decision Diffuser	KUKA	/	✗	-
Zhang et al. (2024a)	Diffuser	CALVIN, CLEVR-Robot	/	✗	-
Ada et al. (2024)	Diffusion-QL	✓	/	✓	-
Ren et al. (2024)	Diffusion-QL	Robomimic, D3IL, FurnitureBench	100-300, 96,50	✓*2	50
Huang et al. (2025b)	Diffusion-QL	MetaWorld, Adroit	20, 50	✓	50
Carvalho et al. (2023)	✗	custom	25	✗	-

For each benchmark, the numbers of demonstrations are listed in the same order. In the column "Diffusion Baselines" only those baselines, which are diffusion methods themselves, are listed. Methods not evaluated against a diffusion-based baseline, indicated by an (✗), are only evaluated against non-diffusion baselines or ablations of the method. The references for the benchmarks are listed in Supplementary Appendix Table 2. In the following, the symbols are explained: *1 As the number refers to the number of transitions, not demonstrations, this high number is expected. A (✓) in the column "Benchmark" indicates that the method is evaluated in simulation, but not with a robotic manipulation task, while a (✗) indicates that the method is not evaluated in simulation. The column "Real" indicates whether methods are evaluated in the real world (✓), or not (✗). A "/" indicates that the information is not provided by the cited paper, while a "-" indicates that the information does not apply.

(Chi et al., 2023), as many methods are developed based on it. A common baseline for methods integrating 3D visual representations is 3D Diffusion Policy (Ze et al., 2024). 3D Diffusion Policy is evaluated against DP, and outperforms it on a huge variety of tasks in the benchmarks Adroit, MetaWorld, and Dexart with an average success rate of 74.4%, outperforming DP by 24.2%. It is also evaluated on four real-world manipulation tasks: rolling and pinching a dumpling, drilling, and pouring. With an average success rate of 85.0% it outperforms DP by 50%. 3D Diffusion Policy is greatly outperformed by 3D Diffuser Actor (Ke et al., 2024) on the CALVIN benchmark, especially for zero-shot long-horizon tasks. However, no comparison for real-world tasks is provided.

The majority of methods are evaluated in simulation as well as in real-world experiments. For real-world experiments, most policies are directly trained on real-world data. However, some are trained exclusively in simulation and applied in the real world in a zero shot (Yu et al., 2023; Mishra et al., 2023; Ren et al., 2024; Liu et al., 2023b; Kapelyukh et al., 2024; Liu et al., 2023c), utilizing domain randomization, or real-world scene reconstruction in simulation. Few, predominately RL methods, are only evaluated in simulation (Yang et al., 2023; Power et al., 2023; Wang et al., 2023a; Janner et al., 2022; Pearce et al., 2022; Wang et al., 2023b; Mendez-Mendez et al., 2023; Kim S. et al., 2024; Brehmer et al., 2023; Liang et al., 2023; Zhou H. et al., 2024; Mishra and Chen, 2024; Ajay et al., 2023; Ding and Jin, 2023; Zhang E. et al., 2024).

6 Conclusion, limitations and outlook

Diffusion models (DMs) have emerged as state-of-the-art methods in robotic manipulation, offering exceptional ability in

TABLE 7 Benchmarks of trajectory diffusion using imitation learning.

Reference	Diffusion Baseline	Simulation		Real World	
		Benchmark	#Demos	Real	#Demos
Diffusion Policy (DP) (Chi et al., 2023)	✗	FrankaKitchen, Robomimic, custom	566,500, /	✓	/
ChainedDiffuser (Xian et al., 2023)	✗	RLBench	100	✓	10 - 20
BESO (Reuss et al., 2023)	DP, Diffusion-BC	Relay Kitchen, CALVIN, custom	566, /,1000	✗	-
Chen et al. (2023a)	✗	CALVIN, FrankaKitchen, Ravens	200K, 566, 1000	✓	/
Zhou et al. (2023)	Diffuser [*1],Decision Diffuser [*2]	RLBench		✗	-
Diffusion-BC (Pearce et al., 2022)	✗	D4RLKitchen	566	✗	-
Mendez-Mendez et al. (2023)	✗	BEHAVIOR	-	✗	-
3D-DP (Ze et al., 2024)	DP	e.g. Adroit, MetaWorld, DexDeform	10 - 100	✓	40
Ke et al. (2024)	3D-DP, ChainedDiffuser	RLBench, CALVIN	24 h	✓	15
Liu et al. (2023c)	DP, SE(3)-DM	Custom	/	✓	/
Power et al. (2023)	✗	custom	/	✗	-
Ma et al. (2024b)	DP, Diffuser	RLBench	100	✓	20
Vosylius et al. (2024)	DP	RLBench	20	✓	20
Zhang et al. (2024a)	Diffuser	CALVIN	/	✗	-
Reuss et al. (2024a)	✗	CALVIN, LIBERO	24 h, 50	✓	4.5 h
Scheikl et al. (2024)	DP, BESO	LapGym	90 - 200	✓	90 - 200
Chen et al. (2024a)	DP, SE(3)-DM	FrankaKitchen, Adroit	16k - 64k, 1.25k - 5 k	✓	60
Zhou et al. (2024a)	DP, BESO, Consistency Models [*3]	Relay Kitchen, XArm Block Push, D3IL	566, 1k,96 - 2 k	✗	-
Li et al. (2024c)	DP, 3D-DP	Robomimic, custom	500,100	✓	100
Si et al. (2024)	DP	✗	-	✓	25 - 50
Saha et al. (2024)	✗	MπNets	6.54Mil	✗	-
Bharadhwaj et al. (2024b)	✗	EpicKitchens, RT1, BridgeData	400 k [*4]	✓	400
Wang et al. (2024b)	✗	custom	50 K trans [*5]	✓	50 K trans [*5]
Li et al. (2025)	DP, 3D-DP	RLBench	40	✓	40
Reuss et al. (2024b)	DP	CALVIN, LIBERO, Relay Kitchen, Block Push	22966, 50, 566, 1000	✗	-

For each benchmark, the numbers of demonstrations are listed in the same order. In the column "Diffusion Baselines" only those baselines, which are diffusion methods themselves, are listed. Methods not evaluated against a diffusion-based baseline, indicated by an (✗), are only evaluated against non-diffusion baselines or ablations of the method. The references for the benchmarks are listed in Supplementary Appendix Table 2. In the following, he symbols are explained: Methods by and [*1] Janner et al. (2022), and [*2] Ajay et al. (2023), and [*3] (Song et al., 2024). [*4] The diffusion model is trained using uncurated video data. [*5] As the number refers to the number of transitions, not demonstrations, this high number is expected. The column "Real" indicates whether methods are evaluated in the real world (✓), or not (✗). A "/" indicates that the information is not provided by the cited paper, while a "-" indicates that the information does not apply.

modeling multi-modal distributions, high training stability, and stability to high-dimensional input and output spaces. Several tasks, challenges, and limitations in the domain of robotic manipulation with DMs remain unsolved. A prevalent issue is the lack of generalizability. The slow inference time for DMs also remains a major bottleneck.

6.1 Limitations

6.1.1 Generalizability

While a lot of methods demonstrate relatively good generalizability in terms of object types, lightning conditions, and task complexity, they still face limitations in this area. This prevalent limitation shared across almost all methodologies in robotic manipulation remains a crucial research question.

The majority of methods using DMs for trajectory generation rely on imitation learning, using mostly behavior cloning. Thus, they inherit the dependence on the quality and diversity of training data, making it difficult to handle out-of-distribution situations due to the covariate shift problem (Ross and Bagnell, 2010). As most methodologies combining DMs with RL use offline RL, they still rely on existing data, mapping a sufficient amount of the state-action space, and are thus also unable to react to distribution shifts. Moreover, offline RL requires more careful fine-tuning than imitation learning to ensure training stability and prevent overfitting. Still, the advantage of RL is that it can handle suboptimal behavior Levine et al. (2020).

While data scaling offers improved generalizability, it typically demands large training datasets and substantial computational resources. One recent solution is to use pre-trained foundation models. Moreover, as the majority of current methods for data augmentation in DMs do not augment trajectories, e.g., (Yu et al., 2023; Mandi et al., 2022), it only increases robustness to slightly different task settings, such as changes in colors, textures, distractors, and background. VLAs can generalize to multi-task and long-horizon settings but often lack action precision, thus requiring finetuning and the combination with more specialized agents (Zhang et al., 2024g).

6.1.2 Sampling speed

The principal limitation inherent to DMs can be attributed to the iterative nature of the sampling process, which results in a time-intensive sampling procedure, thus impeding efficiency and real-time prediction capabilities. Despite recent advances that improve sampling speed and quality (Chen K. et al., 2024; Zhou H. et al., 2024), a considerable number of recent methods use DDIM (Song J. et al., 2021), although other methods, such as DPM-solver (Lu et al., 2022) have shown better performance. However, this comparison has only been performed using image generation benchmarks and would need to be verified for applications in robotic manipulation. Numerous works have demonstrated competitive task performance using DDIM, but do not directly investigate the decrease in task performance associated with a lower number of reverse diffusion steps. Ko et al. (2024) analyzes their approach using both DDPM and DDIM sampling, reporting a sampling process that is ten times faster with only a 5.6% decrease in task performance when using DDIM. Although such a decline might appear negligible, its significance is highly task-dependent. Consequently, there is a need for efficient sampling strategies and a more comprehensive analysis of existing sampling methods, particularly regarding the domain of robotic manipulation. It should, however, be noted that already in DP (Chi et al., 2023), one of the earlier methods combining DMs with receding-horizon control for trajectory planning, real-time control is possible. Using DDIM with 10 denoising steps during inference, they report an inference latency of 0.1 s on a Nvidia 3080 GPU.

6.2 Conclusion and outlook

This survey, to the best to our knowledge, is the first survey reviewing the state-of-the-art methods diffusion models (DMs) in robotics manipulation. This paper offers a thorough discussion of various methodologies regarding network architecture, learning framework, application, and evaluation, highlighting limits and advantages. We explored the three primary applications of DMs in robotic manipulation: trajectory generation, robotic grasping, and visual data augmentation. Most notably, DMs offer exceptional ability in modeling multi-modal distributions, high training stability, and robustness to high-dimensional input and output spaces. Especially in visual robotic manipulation, DMs provide essential capabilities to process high-resolution 2D and 3D visual observations, as well as to predict high-dimensional trajectories and grasp poses, even directly in image space.

A key challenge of DMs is the slow inference speed. In the field of computer vision, fast samplers have been developed that have not yet been evaluated in the field of robotic manipulation. Testing those samplers and comparing them against the commonly used ones, could be one step to increase sampling efficiency. Moreover, there are also methods for fast sampling, specifically in robotic manipulation, that are not broadly used, e.g. BRIDGeR (Chen K. et al., 2024). While the generalizability of DMs remains also an open challenge, the image generation capabilities of DMs open new avenues in data augmentation for data scaling, making methods more robust to limited data variety. Generalizability could be also improved by the integration of advanced vision-language, and vision-language action models.

We believe continual learning could be a promising approach to improve generalizability and adaptability in highly dynamic and unfamiliar environments. This remains a widely unexplored problem domain for DMs in robotic manipulation, exceptions are (Di Palo et al., 2024; Mendez-Mendez et al., 2023). However, these methods have strong limitations. For instance, Di Palo et al. (2024) relies on precise feature descriptions of all involved objects and is restricted to predefined abstract skills. Moreover, their continual update process involves replaying all past data, which is both computationally inefficient and does not prevent catastrophic forgetting. Morover, to handle complex and cluttered scenes, view planning and iterative planning strategies, also considering complete occlusions, could be combined with existing DMs using 3D scene representations. Leveraging the semantic reasoning capabilities of vision language and vision language action models could be a possible approach.

Author contributions

RW: Writing – original draft, Writing – review and editing. YS: Writing – review and editing, Writing – original draft. SL: Writing – original draft. RR: Writing – original draft, Supervision, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB-1574 – 471687386.

Acknowledgments

We thank our colleague Edgar Welte for providing the video data for the illustration of the diffusion process in Figure 2.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2025.1606247/full#supplementary-material

References

Ada, S. E., Oztop, E., and Ugur, E. (2024). Diffusion policies for out-of-distribution generalization in offline reinforcement learning. *IEEE Robotics Automation Lett.* 9, 3116–3123. doi:10.1109/LRA.2024.3363530

Ajay, A., Du, Y., Gupta, A., Tenenbaum, J., Jaakkola, T., and Agrawal, P. (2023). "IS conditional generative modeling all you need for decision-making?," in The Eleventh International Conference on Learning Representations.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., et al. (2017). Hindsight experience replay. *Adv. Neural Inf. Process. Syst.* 30. Available online at: https://proceedings.neurips.cc/paper_files/paper/2017.

Barad, K. R., Orsula, A., Richard, A., Dentler, J., Olivares-Mendez, M. A., and Martinez, C. (2024). GraspLDM: generative 6-DoF grasp synthesis using latent diffusion models. *IEEE Access* 12, 164621–164633. doi:10.1109/ACCESS.2024.3492118

Bharadhwaj, H., Gupta, A., Kumar, V., and Tulsiani, S. (2024a). "Towards generalizable zero-shot manipulation via translating human interaction plans," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 6904–6911. doi:10.1109/ICRA57147.2024.10610288

Bharadhwaj, H., Mottaghi, R., Gupta, A., and Tulsiani, S. (2024b). "Track2Act: predicting point tracks from internet videos enables generalizable robot manipulation," in 1st Workshop on X-Embodiment Robot Learning, 306–324. doi:10.1007/978-3-031-73116-7_18

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., et al. (2024a). *π_0: a vision-language-action flow model for general robot control*. arXiv preprint arXiv:2410.24164.

Black, K., Nakamoto, M., Atreya, P., Walke, H., Finn, C., Kumar, A., et al. (2024b). "Zero-shot robotic manipulation with pretrained image-editing diffusion models," in 12th International Conference on Learning Representations, ICLR 2024.

Bohg, J., Morales, A., Asfour, T., and Kragic, D. (2013). Data-driven grasp synthesis—a survey. *IEEE Trans. robotics* 30, 289–309. doi:10.1109/tro.2013.2289018

Braun, M., Jaquier, N., Rozo, L., and Asfour, T. (2024). "Riemannian flow matching policy for robot motion learning," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 5144–5151. doi:10.1109/IROS58592.2024.10801521

Brehmer, J., Bose, J., de Haan, P., and Cohen, T. S. (2023). EDGI: equivariant diffusion for planning with embodied agents. *Adv. Neural Inf. Process. Syst.* 36, 63818–63834. Available online at: https://proceedings.neurips.cc/paper_files/paper/2023.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., et al. (2023a). *RT-2: vision-Language-action models transfer web knowledge to robotic control*. arXiv preprint arXiv:2307.15818.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., et al. (2023b). *RT-1: robotics transformer for real-world control at scale*. arXiv preprint arXiv:2212.06817.

Cao, J., Liu, J., Kitani, K., and Zhou, Y. (2024). *Multi-modal diffusion for hand-object grasp generation*. arXiv preprint arXiv:2409.04560.

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 International Conference on Intelligent Robots and Systems, 1916–1923. doi:10.1109/IROS55552.2023.10342382

Carvalho, J., Le, A. T., Jahr, P., Sun, Q., Urain, J., Koert, D., et al. (2024). *Grasp diffusion network: learning grasp generators from partial point clouds with diffusion models in SO (3) xR3*. arXiv preprint arXiv:2412.08398.

Chang, X., and Sun, Y. (2024). *Text2Grasp: grasp synthesis by text prompts of object grasping parts*. arXiv preprint arXiv:2404.15189.

Cheang, C.-L., Chen, G., Jing, Y., Kong, T., Li, H., Li, Y., et al. (2024). *GR-2: a generative video-language-action model with web-scale knowledge for robot manipulation*. arXiv preprint arXiv:2410.06158.

Chen, K., Lim, E., Lin, K., Chen, Y., and Soh, H. (2024a). Don't start from scratch: behavioral refinement via interpolant-based policy diffusion. *Robotics Sci. Syst.* doi:10.48550/arXiv.2402.16075

Chen, L., Bahl, S., and Pathak, D. (2023a). PlayFusion: Skill acquisition via diffusion from language-annotated play. *Proc. 7th Conf. Robot Learn.* 229, 2012–2029. Available online at: https://proceedings.mlr.press/v229/chen23c.html.

Chen, L. Y., Xu, C., Dharmarajan, K., Irshad, M. Z., Cheng, R., Keutzer, K., et al. (2024b). "Rovi-aug: robot and viewpoint augmentation for cross-embodiment robot learning," in Conference on Robot Learning (CoRL).

Chen, Z., Kiami, S., Gupta, A., and Kumar, V. (2023b). *GenAug: retargeting behaviors to unseen situations via generative augmentation*. arXiv preprint arXiv:2302.06671.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., et al. (2023). Diffusion policy: visuomotor policy learning via action diffusion. *Robotics Sci. Syst. (RSS)*. doi:10.48550/arXiv.2303.04137

Dhariwal, P., and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Adv. Neural Inf. Process. Syst.* 34, 8780–8794. Available online at: https://proceedings.neurips.cc/paper_files/paper/2021.

Ding, Z., and Jin, C. (2023). "Consistency models as a rich and efficient policy class for reinforcement learning," in International Conference on Robot Learning.

Di Palo, N., Hasenclever, L., Humplik, J., and Byravan, A. (2024). *Diffusion augmented agents: a framework for efficient exploration and transfer learning*. arXiv preprint arXiv:2407.20798.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16x16 words: transformers for image recognition at scale," in International Conference on Learning Representations.

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., et al. (2023). Learning universal policies via text-guided video generation. *Adv. Neural Inf. Process. Syst.* 36, 9156–9172. Available online at: https://proceedings.neurips.cc/paper_files/paper/2023

Fang, H.-S., Wang, C., Gou, M., and Lu, C. (2020). "Graspnet-1billion: a large-scale benchmark for general object grasping," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 11444–11453.

Feng, Q., Feng, J., Chen, Z., Triebel, R., and Knoll, A. (2024). *FFHFlow: a flow-based variational approach for multi-fingered grasp synthesis in real time*. arXiv preprint arXiv:2407.15161.

Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., et al. (2025). Foundation models in robotics: applications, challenges, and the future. *Int. J. Robotics Res.* 44, 701–739. doi:10.1177/02783649241281508

Florence, P., Lynch, C., Zeng, A., Ramirez, O., Wahid, A., Downs, L., et al. (2022). Implicit behavioral cloning. *Proc. Mach. Learn. Res.* 164, 158–168. Available online at: https://proceedings.mlr.press/v164/florence22a.

Frans, K., Hafner, D., Levine, S., and Abbeel, P. (2025). "One step diffusion via shortcut models," in The Thirteenth International Conference on Learning Representations.

Freiberg, R., Qualmann, A., Vien, N. A., and Neumann, G. (2025). Diffusion for multi-embodiment grasping. *IEEE Robotics Automation Lett.* 10, 2694–2701. doi:10.1109/LRA.2025.3534065

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). *D4RL: datasets for deep data-driven reinforcement learning*. arXiv preprint arXiv:2004.07219.

Geng, J., Liang, X., Wang, H., and Zhao, Y. (2023). "Diffusion policies as multi-agent reinforcement learning strategies," in *Lecture notes in computer science*, 356–364.

Gervet, T., Xian, Z., Gkanatsios, N., and Fragkiadaki, K. (2023). Act3D: 3D feature field transformers for multi-task robotic manipulation. *Proc. 7th Conf. Robot Learn.* 229, 3949–3965. Available online at: https://proceedings.mlr.press/v229/gervet23a.html.

Gilles, M., Chen, Y., Zeng, E. Z., Wu, Y., Furmans, K., Wong, A., et al. (2023). Metagraspnetv2: all-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping. *IEEE Trans. Automation Sci. Eng.* 21, 2302–2320. doi:10.1109/tase.2023.3328964

Gilles, M., Furmans, K., and Rayyes, R. (2025). MetaMVUC: active learning for sample-efficient sim-to-real domain adaptation in robotic grasping. *IEEE Robotics Automation Lett.* 10, 3644–3651. doi:10.1109/LRA.2025.3544083

Goyal, A., Xu, J., Guo, Y., Blukis, V., Chao, Y.-W., and Fox, D. (2023). "Rvt: robotic view transformer for 3d object manipulation," in Conference on Robot Learning, 694–710.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., et al. (2022). "Vector quantized diffusion model for text-to-image synthesis," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10686–10696. doi:10.1109/CVPR52688.2022.01043

Gupta, A., Kumar, V., Lynch, C., Levine, S., and Hausman, K. (2020). Relay policy learning: solving long-horizon tasks via imitation and reinforcement learning. *Proc. Mach. Learn. Res.*, 1025–1037. Available online at: https://proceedings.mlr.press/v100/gupta20a.

Ha, H., Florence, P., and Song, S. (2023). Scaling up and distilling Down: language-guided robot skill acquisition. *Proc. 7th Conf. Robot Learn.* 229, 3766–3777. Available online at: https://proceedings.mlr.press/v229/ha23a.html.

Ho, J., and Ermon, S. (2016). "Generative adversarial imitation learning," in *Advances in neural information processing systems*.

Ho, J., Jain, A., and Abbeel, P. (2020). "Denoising diffusion probabilistic models," in Proceedings of the 34th International Conference on Neural Information Processing Systems.

Ho, J., Research, G., and Salimans, T. (2021). "Classifier-free diffusion guidance," in NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

Huang, D., Dong, W., Tang, C., and Zhang, H. (2025a). *HGDiffuser: efficient task-oriented grasp generation via human-guided grasp diffusion models*. arXiv preprint arXiv:2503.00508.

Huang, H., Wang, D., Zhu, X., Walters, R., and Platt, R. (2023). "Edge grasp network: a graph-based se (3)-invariant approach to grasp detection," in 2023 IEEE International Conference on Robotics and Automation (ICRA), 3882–3888. doi:10.1109/icra48891.2023.10160728

Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., et al. (2024a). "An embodied generalist agent in 3D world," in Proceedings of the 41st International Conference on Machine Learning.

Huang, T., Jiang, G., Ze, Y., Xu, H., Qi, S., and Institute, Z. (2025b). "Diffusion reward: learning rewards via conditional video diffusion," in Computer Vision – ECCV, 478–495. doi:10.1007/978-3-031-72946-1_27

Huang, Z., Lin, Y., Yang, F., and Berenson, D. (2024b). "Subgoal diffuser: coarse-to-fine subgoal generation to guide model predictive control for robot manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 16489–16495. doi:10.1109/ICRA57147.2024.10610189

Iioka, Y., Yoshida, Y., Wada, Y., Hatanaka, S., and Sugiura, K. (2023). "Multimodal diffusion segmentation model for object segmentation from manipulation instructions," in IEEE International Conference on Intelligent Robots and Systems, 7590–7597. doi:10.1109/IROS55552.2023.10341402

Ikeda, T., Zakharov, S., Ko, T., Irshad, M. Z., Lee, R., Liu, K., et al. (2024). "Diffusionnocs: managing symmetry and uncertainty in sim2real multi-modal category-level pose estimation," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 7406–7413. doi:10.1109/IROS58592.2024.10802487

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. (2020). RLBench: the robot learning benchmark & learning environment. *IEEE Robotics Automation Lett.* 5, 3019–3026. doi:10.1109/LRA.2020.2974707

Janner, M., Du, Y., Tenenbaum, J., and Levine, S. (2022). Planning with diffusion for flexible behavior synthesis. *Proc. 39th Int. Conf. Mach. Learn.* 162, 9902–9915. Available online at: https://proceedings.mlr.press/v162/janner22a.html.

Jiang, Z., Zhu, Y., Svetlik, M., Fang, K., and Zhu, Y. (2021). *Synergies between affordance and geometry: 6-dof grasp detection via implicit representations*. arXiv preprint arXiv:2104.01542.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., and Kachman, T. (2021). "Gotta Go fast with score-based generative models," in The Symposium of Deep Learning and Differential Equations.

Kang, B., Ma, X., Du, C., Pang, T., and Yan, S. (2023). "Efficient diffusion policies for offline reinforcement learning," in *Advances in neural information processing systems*, 67195–67212.

Kapelyukh, I., Ren, Y., Alzugaray, I., and Johns, E. (2024). "Dream2Real: zero-shot 3D object rearrangement with vision-language models," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 4796–4803. doi:10.1109/ICRA57147.2024.10611220

Kapelyukh, I., Vosylius, V., and Johns, E. (2023). DALL-E-Bot: introducing web-scale diffusion models to robotics. *IEEE Robotics Automation Lett.* 8, 3956–3963. doi:10.1109/LRA.2023.3272516

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* 35, 26565–26577. Available online at: https://proceedings.neurips.cc/paper_files/paper/2022.

Kasahara, I., Agrawal, S., Engin, S., Chavan-Dafle, N., Song, S., and Isler, V. (2024). "RIC: rotate-inpaint-complete for generalizable scene reconstruction," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2713–2720. doi:10.1109/ICRA57147.2024.10611694

Katara, P., Xian, Z., and Fragkiadaki, K. (2024). "Gen2Sim: scaling up robot learning in simulation with generative models," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 6672–6679. doi:10.1109/ICRA57147.2024.10610566

Ke, T.-W., Gkanatsios, N., and Fragkiadaki, K. (2024). "3D diffuser actor: policy diffusion with 3D scene representations," in 8th Annual Conference on Robot Learning.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., et al. (2024a). "OpenVLA: an open-source vision-language-action model," in 8th Annual Conference on Robot Learning.

Kim, S., Choi, Y., Matsunaga, D. E., and Kim, K.-E. (2024b). Stitching sub-trajectories with conditional diffusion model for goal-conditioned offline RL. *Proc. AAAI Conf. Artif. Intell.* 38, 13160–13167. doi:10.1609/aaai.v38i12.29215

Kim, W. K., Yoo, M., and Woo, H. (2024c). Robust policy learning via offline skill diffusion. *Proc. AAAI Conf. Artif. Intell.* 38, 13177–13184. doi:10.1609/aaai.v38i12.29217

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4015–4026.

Ko, P.-C., Mao, J., Du, Y., Sun, S.-H., and Tenenbaum, J. B. (2024). "Learning to act from actionless videos through dense correspondences," in The Twelth Internactional Conference on Learning Representations.

Krichen, M. (2023). "Generative adversarial networks," in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–7. doi:10.1109/ICCCNT56998.2023.10306417

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). *Offline reinforcement learning: tutorial, review, and perspectives on open problems*. CoRR abs/2005.01643.

Li, H., Feng, Q., Zheng, Z., Feng, J., Chen, Z., and Knoll, A. (2025). *Language-guided object-centric diffusion policy for generalizable and collision-aware robotic manipulation*. arXiv preprint arXiv:2407.00451.

Li, P., Wang, Z., Liu, M., Liu, H., and Chen, C. (2024a). "ClickDiff: click to induce semantic contact map for controllable grasp generation with diffusion models," in Proceedings of the 32nd ACM International Conference on Multimedia, 273–281. doi:10.1145/3664647.3680597

Li, Q., Liang, Y., Wang, Z., Luo, L., Chen, X., Liao, M., et al. (2024b). *CogACT: a foundational vision-language-action model for synergizing cognition and action in robotic manipulation*. arXiv preprint arXiv:2411.19650.

Li, W., Wang, X., Jin, B., and Zha, H. (2023). Hierarchical diffusion for offline decision making. *Proc. 40th Int. Conf. Mach. Learn.* 202, 20035–20064. Available online at: https://proceedings.mlr.press/v202/li23ad.html.

Li, X., Belagali, V., Shang, J., and Ryoo, M. S. (2024c). "Crossway diffusion: improving diffusion-based visuomotor policy via self-supervised learning," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 16841–16849. doi:10.1109/ICRA57147.2024.10610175

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. (2022). Diffusion-LM improves controllable text generation. *Adv. Neural Inf. Process. Syst.* 35, 4328–4343. Available online at: https://proceedings.neurips.cc/paper_files/paper/2022.

Li, Y., Wu, Z., Zhao, H., Yang, T., Liu, Z., Shu, P., et al. (2024d). *ALDM-Grasping: diffusion-aided zero-shot sim-to-real transfer for robot grasping*. arXiv preprint arXiv:2403.11459.

Liang, Z., Mu, Y., Ding, M., Ni, F., Tomizuka, M., and Luo, P. (2023). AdaptDiffuser: diffusion models as adaptive self-evolving planners. *Proc. 40th Int. Conf. Mach. Learn.* 202, 20725–20745. Available online at: https://proceedings.mlr.press/v202/liang23e.html.

Liang, Z., Mu, Y., Ma, H., Tomizuka, M., Ding, M., and Luo, P. (2024). "SkillDiffuser: interpretable hierarchical planning via skill abstractions in diffusion-based task execution," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16467–16476. doi:10.1109/cvpr52733.2024.01558

Lim, B., Kim, J., Kim, J., Lee, Y., and Park, F. C. (2024). "EquiGraspFlow: SE (3)-Equivariant 6-DoF grasp pose generative flows," in 8th Annual Conference on Robot Learning.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). "Flow matching for generative modeling," in The Eleventh International Conference on Learning Representations.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., et al. (2023a). LIBERO: benchmarking knowledge transfer for lifelong robot learning. *Adv. Neural Inf. Process. Syst.* 36, 44776–44791. Available online at: https://proceedings.neurips.cc/paper_files/paper/2023.

Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., et al. (2024). *RDT-1B: a diffusion foundation model for bimanual manipulation*. arXiv preprint arXiv:2410.07864.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., et al. (2025). "Grounding dino: marrying dino with grounded pre-training for open-set object detection," in Computer Vision – ECCV 2024, 38–55. doi:10.1007/978-3-031-72970-6_3

Liu, W., Du, Y., Hermans, T., Chernova, S., and Paxton, C. (2023b). StructDiffusion: language-guided creation of physically-valid structures using unseen objects. *Robotics Sci. Syst*. doi:10.15607/RSS.2023.XIX.031

Liu, W., Mao, J., Hsu, J., Hermans, T., Garg, A., and Wu, J. (2023c). Composable part-based manipulation. *Proc. 7th Conf. Robot Learn.* 229, 1300–1315. Available online at: https://proceedings.mlr.press/v229/liu23e.html.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022). "DPM-Solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Advances in neural information processing systems*, 5775–5787.

Lu, J., Kang, H., Li, H., Liu, B., Yang, Y., Huang, Q., et al. (2025). "Ugg: unified generative grasping," in Computer Vision – ECCV 2024, 414–433. doi:10.1007/978-3-031-72855-6_24

Lucic, M., Kurach, K., Google, M. M., Bousquet, B. O., and Gelly, S. (2018). Are GANs created equal? A large-scale study. *Adv. Neural Inf. Process. Syst.* 30. Available online at: https://proceedings.neurips.cc/paper_files/paper/2018.

Ma, C., Yang, H., Zhang, H., Liu, Z., Zhao, C., Tang, J., et al. (2024a). *DexDiff: towards extrinsic dexterity manipulation of ungraspable objects in unrestricted environments*. arXiv preprint arXiv:2409.05493.

Ma, X., Patidar, S., Haughton, I., and James, S. (2024b). "Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18081–18090. doi:10.1109/cvpr52733.2024.01712

Mandi, Z., Bharadhwaj, H., Moens, V., Song, S., Rajeswaran, A., and Kumar, V. (2022). "CACTI: a framework for scalable multi-task multi-scene visual imitation learning," in CoRL 2022 Workshop on Pre-training Robot Learning.

Maria Scheikl, P., Gyenes, B., Younis, R., Haas, C., Neumann, G., Wagner, M., et al. (2023). LapGym-An open source framework for reinforcement learning in robot-assisted laparoscopic surgery. *J. Mach. Learn. Res.* 24, 1–42. Available online at: http://jmlr.org/papers/v24/23-0207.html.

Martinez, F., Jacinto, E., and Montiel, H. (2023). Rapidly exploring random trees for autonomous navigation in observable and uncertain environments. *Int. J. Adv. Comput. Sci. Appl.* 14. doi:10.14569/IJACSA.2023.0140399

Mattingley, J., Wang, Y., and Boyd, S. (2011). Receding horizon control. *IEEE Control Syst. Mag.* 31, 52–65. doi:10.1109/MCS.2011.940571

Mees, O., Hermann, L., Rosete-Beas, E., and Burgard, W. B. (2022). CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics Automation Lett.* 7, 7327–7334. doi:10.1109/LRA.2022.3180108

Meila, M., and Zhang, T. (2021). "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning (PMLR), 8748–8763. Available online at: https://proceedings.mlr.press/v139/radford21a.

Mendez-Mendez, J., Kaelbling, L. P., and Lozano-Pérez, T. (2023). Embodied lifelong learning for task and motion planning. *Proc. 7th Conf. Robot Learn.* 229, 2134–2150. Available online at: https://proceedings.mlr.press/v229/mendez-mendez23a.html.

Meyer-Veit, F., Rayyes, R., Gerstner, A. O., and Steil, J. (2022a). *Hyperspectral wavelength analysis with u-net for larynx cancer detection*. Cham: Springer Nature Switzerland.

Meyer-Veit, F., Rayyes, R., Gerstner, A. O. H., and Steil, J. (2022b). Hyperspectral endoscopy using deep learning for laryngeal cancer segmentation. *Artif. Neural Netw. Mach. Learn. – ICANN* 2022, 682–694. doi:10.1007/978-3-031-15937-4_57

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). "NeRF: representing scenes as neural radiance fields for view synthesis," in Computer Vision – ECCV 2020: 16th European Conference Proceedings, Part I, Glasgow, UK, August 23–28, 2020, 405–421. doi:10.1007/978-3-030-58452-8_24

Mishra, U. A., and Chen, Y. (2024). "ReorientDiff: diffusion model based reorientation for object manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 10867–10873. doi:10.1109/ICRA57147.2024.10610749

Mishra, U. A., Xue, S., Chen, Y., and Xu, D. (2023). Generative skill chaining: long-horizon skill planning with diffusion models. *Proc. 7th Conf. Robot Learn.* 229, 2905–2925. Available online at: https://proceedings.mlr.press/v229/mishra23a.html.

Misra, D. (2019). *Mish: a self regularized non-monotonic activation function*. arXiv preprint arXiv:1908.08681.

Mousavian, A., Eppner, C., and Fox, D. (2019). "6-DOF GraspNet: variational grasp generation for object manipulation," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2901–2910. doi:10.1109/iccv.2019.00299

Newbury, R., Gu, M., Chumbley, L., Mousavian, A., Eppner, C., Leitner, J., et al. (2023). Deep learning approaches to grasp synthesis: a review. *IEEE Trans. Robotics* 39, 3994–4015. doi:10.1109/tro.2023.3280597

Nguyen, K., Le, A. T., Pham, T., Huber, M., Peters, J., and Vu, M. N. (2025). *FlowMP: learning motion fields for robot planning with conditional flow matching*. arXiv preprint arXiv:2503.06135.

Nguyen, N., Vu, M. N., Huang, B., Vuong, A., Le, N., Vo, T., et al. (2024a). "Lightweight language-driven grasp detection using conditional consistency model," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 13719–13725doi. doi:10.1109/IROS58592.2024.10802007

Nguyen, T., Vu, M. N., Huang, B., Van Vo, T., Truong, V., Le, N., et al. (2024b). "Language-conditioned affordance-pose detection in 3D point clouds," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 3071–3078. doi:10.1109/ICRA57147.2024.10610008

Nichol, A. Q., and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *Proc. 38th Int. Conf. Mach. Learn.* 139, 8162–8171. Available online at: https://proceedings.mlr.press/v139/nichol21a.html.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al. (2023). *Dinov2: learning robust visual features without supervision*. arXiv preprint arXiv:2304.07193.

Pan, C., Junge, K., and Hughes, J. (2024a). *Vision-language-action model and diffusion policy switching enables dexterous control of an anthropomorphic hand*. arXiv preprint arXiv:2410.14022.

Pan, S., Jin, L., Huang, X., Stachniss, C., Popović, M., and Bennewitz, M. (2024b). "Exploiting priors from 3D diffusion models for RGB-based one-shot view planning," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 13341–13348. doi:10.1109/IROS58592.2024.10802551

Pan, S., Jin, L., Huang, X., Stachniss, C., Popović, M., and Bennewitz, M. (2025). *Dm-osvp++: one-shot view planning using 3d diffusion models for active rgb-based object reconstruction*. arXiv preprint arXiv:2504.11674.

Pearce, T., Rashid, T., Kanervisto, A., Bignell, D., Sun, M., Georgescu, R., et al. (2022). "Imitating human behaviour with diffusion models," in Deep Reinforcement Learning Workshop NeurIPS 2022. doi:10.48550/arXiv.2301.10677

Peebles, W., and Xie, S. (2023). "Scalable diffusion models with transformers," in Proceedings of the IEEE International Conference on Computer Vision, 4172–4182. doi:10.1109/ICCV51070.2023.00387

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). *FiLM: visual reasoning with a general conditioning layer*. Proceedings of the AAAI Conference on Artificial Intelligence 32, (1). doi:10.1609/aaai.v32i1.11671

Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., et al. (2025). *FAST: efficient action tokenization for vision-language-action models*. arXiv preprint arXiv:2501.0974.

Pfrommer, D., Padmanabhan, S., Ahn, K., Umenberger, J., Marcucci, T., Mhammedi, Z., et al. (2024). "On the sample complexity of imitation learning for smoothed model predictive control," in 2024 IEEE 63rd Conference on Decision and Control (CDC), 1820–1825doi. doi:10.1109/CDC56724.2024.10886242

Power, T., Soltani-Zarrin, R., Iba, S., and Berenson, D. (2023). "Sampling constrained trajectories using composable diffusion models," in IROS 2023 Workshop on Differentiable Probabilistic Robotics: Emerging Perspectives on Robot Learning.

Prasad, A., Lin, K., Wu, J., Zhou, L., and Bohg, J. (2024). Consistency policy accelerated visuomotor policies via consistency distillation. *Robotics Sci. Syst.* doi:10.48550/arXiv.2405.07503

Qi, C., Haramati, D., Daniel, T., Tamar, A., and Zhang, A. (2025). *Ec-diffuser: multi-object manipulation via entity-centric behavior generation*. arXiv prepping arXiv:2412.18907.

Qian, Y., Zhu, X., Biza, O., Jiang, S., Zhao, L., Huang, H., et al. (2024). "ThinkGrasp: a vision-language system for strategic part grasping in clutter," in 8th Annual Conference on Robot Learning.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in International conference on machine learning, 8748–8763.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., et al. (2017). *Learning complex dexterous manipulation with deep reinforcement learning and demonstrations*. arXiv preprint arXiv:1709.10087.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents*. arXiv preprint arXiv:2204.06125.

Ren, A. Z., Lidard, J., Ankile, L. L., Simeonov, A., Agrawal, P., Majumdar, A., et al. (2024). "Diffusion policy policy optimization," in CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data.

Reuss, M., Erdinç, O., Gmurlu, Y., Wenzel, F., and Lioutikov, R. (2024a). Multimodal diffusion transformer: learning versatile behavior from multimodal goals. *Robotics Sci. Syst.* doi:10.48550/arXiv.2407.05996

Reuss, M., Li, M., and Lioutikov, R. (2023). Goal-conditioned imitation learning using score-based diffusion policies. *Robotics Sci. Syst.* doi:10.48550/arXiv.2304.02532

Reuss, M., Pari, J., Agrawal, P., and Lioutikov, R. (2024b). *Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning*. arXiv preprint arXiv:2412.12953.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022a). "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022b). "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684–10695.

Römer, R., von Rohr, A., and Schoellig, A. P. (2024). *Diffusion predictive control with constraints*. arXiv preprint/arXiv.2412.09342.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Med. Image Comput. Computer-Assisted Intervention – MICCAI* 2015, 234–241. doi:10.1007/978-3-319-24574-4_28

Ross, S., and Bagnell, D. (2010). Efficient reductions for imitation learning. *Proc. Thirteen. Int. Conf. Artif. Intell. Statistics* 9, 661–668. Available online at: https://proceedings.mlr.press/v9/ross10a.html.

Rouxel, Q., Ferrari, A., Ivaldi, S., and Mouret, J.-B. (2024). "Flow matching imitation learning for multi-support manipulation," in 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids), 528–535. doi:10.1109/Humanoids58906.2024.10769838

Ryu, H., Kim, J., An, H., Chang, J., Seo, J., Kim, T., et al. (2024). "Diffusion-EDFs: Bi-equivariant denoising generative modeling on SE(3) for visual robotic manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18007–18018. doi:10.1109/cvpr52733.2024.01705

Ryu, H., Lee, H., Lee, J.-H., and Choi, J. (2023). "Equivariant descriptor fields: SE(3)-equivariant energy-based models for end-to-end visual robotic manipulation learning," in The Eleventh International Conference on Learning Representations.

Saha, K., Mandadi, V., Reddy, J., Srikanth, A., Agarwal, A., Sen, B., et al. (2024). "EDMP: ensemble-of-costs-guided diffusion for motion planning," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 10351–10358. doi:10.1109/ICRA57147.2024.10610519

Salimans, T., and Ho, J. (2022). "Progressive distillation for fast sampling of diffusion models," in International Conference on Learning Representations (ICLR).

Scheikl, P. M., Schreiber, N., Haas, C., Freymuth, N., Neumann, G., Lioutikov, R., et al. (2024). Movement primitive diffusion: learning gentle robotic manipulation of deformable objects. *IEEE Robotics Automation Lett.* 9, 5338–5345. doi:10.1109/LRA.2024.3382529

Seo, J., Yoo, S., Chang, J., An, H., Ryu, H., Lee, S., et al. (2025). *SE (3)-Equivariant robot learning and control: a tutorial survey*. arXiv preprint arXiv:2503.09829.

Shentu, Y., Wu, P., Rajeswaran, A., and Abbeel, P. (2024). "From LLMs to actions: latent codes as bridges in hierarchical robot control," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 8539–8546. doi:10.1109/IROS58592.2024.10801683

Shi, L. X., Sharma, A., Zhao, T. Z., and Finn, C. (2023). Waypoint-based imitation learning for robotic manipulation. *Proc. 7th Conf. Robot Learn.* 229, 2195–2209. Available online at: https://proceedings.mlr.press/v229/shi23b.html.

Shi, Y., Welte, E., Gilles, M., and Rayyes, R. (2024). *vMF-Contact: uncertainty-aware evidential learning for probabilistic contact-grasp in noisy clutter*. arXiv preprint arXiv:2411.03591.

Shi, Y., Wen, D., Chen, G., Welte, E., Liu, S., Peng, K., et al. (2025). *VISO-Grasp: vision-language informed spatial object-centric 6-DoF active view planning and grasping in clutter and invisibility*. arXiv preprint arXiv:2503.12609.

Si, Z., Zhang, K., Temel, Z., and Kroemer, O. (2024). Tilde: teleoperation for dexterous In-Hand manipulation learning with a DeltaHand. *Robotics Sci. Syst.*

Simeonov, A., Goyal, A., Manuelli, L., Yen-Chen, L., Sarmiento, A., Rodriguez, A., et al. (2023). "Shelving, stacking, hanging: relational pose diffusion for multi-modal rearrangement," in Conference on Robot Learning.

Singh, G., Kalwar, S., Karim, M. F., Sen, B., Govindan, N., Sridhar, S., et al. (2024). "Constrained 6-DoF grasp generation on complex shapes for improved dual-arm manipulation," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 7344–7350. doi:10.1109/IROS58592.2024.10802268

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *Proc. 32nd Int. Conf. Mach. Learn.* 37, 2256–2265. Available online at: https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Song, J., Meng, C., and Ermon, S. (2021a). "Denoising diffusion implicit models," in International Conference on Learning Representations.

Song, P., Li, P., and Detry, R. (2024). "Implicit grasp diffusion: bridging the gap between dense prediction and sampling-based grasping," in 8th Annual Conference on Robot Learning.

Song, Y., and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* 32.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021b). "Score-based generative modeling through stochastic differential equations," in International Conference on Learning Representations.

Suh, H. T., Chou, G., Dai, H., Yang, L., Gupta, A., and Tedrake, R. (2023). Fighting uncertainty with gradients: offline reinforcement learning via diffusion score matching. *Proc. 7th Conf. Robot Learn.* 229, 2878–2904. Available online at: https://proceedings.mlr.press/v229/suh23a.html.

Tarvainen, A., and Valpola, H. (2017). Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* 30. Available online at: https://proceedings.neurips.cc/paper_files/paper/2017.

Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., et al. (2024). *Octo: an open-source generalist robot policy*. arXiv preprint arXiv:2405.12213.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 23–30. doi:10.1109/IROS.2017.8202133

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., et al. (2018). "Training deep networks with synthetic data: bridging the reality gap by domain randomization," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1082–10828. doi:10.1109/CVPRW.2018.00143

Tsagkas, N., Rome, J., Ramamoorthy, S., Aodha, O. M., and Lu, C. X. (2024). "Click to grasp: zero-shot precise manipulation via visual diffusion descriptors," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 11610–11617. doi:10.1109/IROS58592.2024.10801488

Urain, J., Funk, N., Peters, J., and Chalvatzaki, G. (2023). "SE(3)-DiffusionFields: learning smooth cost functions for joint grasp and motion optimization through diffusion," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 5923–5930. doi:10.1109/ICRA48891.2023.10161569

Venkatraman, S., Khaitan, S., Akella, R. T., Dolan, J., Schneider, J., and Berseth, G. (2023). *Reasoning with latent diffusion in offline reinforcement learning*. arXiv preprint arXiv:2309.06599.

Vosylius, V., Seo, Y., Uruç, J., and James, S. (2024). Render and diffuse: aligning image and action spaces for diffusion-based behaviour cloning. *Robotics Sci. Syst.* doi:10.15607/RSS.2024.XX.051

Vuong, A. D., Vu, M. N., Huang, B., Nguyen, N., Le, H., Vo, T., et al. (2024). "Language-driven grasp detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 17902–17912. doi:10.1109/cvpr52733.2024.01695

Wang, C., Shi, H., Wang, W., Zhang, R., Fei-Fei, L., and Liu, C. K. (2024a). "DexCap: scalable and portable mocap data collection system for dexterous manipulation," in 2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS).

Wang, L., Zhao, J., Du, Y., Adelson, E. H., and Tedrake, R. (2024b). PoCo: policy composition from and for heterogeneous robot learning. *Robotics Sci. Syst.* doi:10.48550/arXiv.2402.02511

Wang, Y.-K., Xing, C., Wei, Y.-L., Wu, X.-M., and Zheng, W.-S. (2024c). "Single-view scene point cloud human grasp generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 831–841. doi:10.1109/cvpr52733.2024.00085

Wang, Z., Hunt, J. J., and Zhou, M. (2023a). *Diffusion policies as an expressive policy class for offline reinforcement learning*. arXiv preprint arXiv:2208.06193.

Wang, Z., Oba, T., Yoneda, T., Shen, R., Walter, M., and Stadie, B. C. (2023b). Cold diffusion on the replay buffer: learning to plan from known good States. *Proc. 7th Conf. Robot Learn.* 229, 3277–3291. Available online at: https://proceedings.mlr.press/v229/wang23e.html.

Watson, D., Chan, W., Ho, J., and Norouzi, M. (2022). "Learning fast samplers for diffusion models by differentiating trough sample quality," in International Conference on Learning Representations (ICLR).

Welte, E., and Rayyes, R. (2025). *Interactive imitation learning for dexterous robotic manipulation: challenges and perspectives – a survey*. arXiv prepint arXiv:2506.00098.

Wen, J., Zhu, M., Zhu, Y., Tang, Z., Li, J., Zhou, Z., et al. (2024). *Diffusion-VLA: scaling robot foundation models via unified diffusion and autoregression*. arXiv preprint arXiv:2412.03293.

Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., et al. (2025). TinyVLA: toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics Automation Lett.* 10, 3988–3995. doi:10.1109/LRA.2025.3544909

Weng, Z., Lu, H., Kragic, D., and Lundell, J. (2024). DexDiffuser: generating dexterous grasps with diffusion models. *IEEE Robotics Automation Lett.* 9, 11834–11840. doi:10.1109/LRA.2024.3498776

Wu, T., Gan, Y., Wu, M., Cheng, J., Yang, Y., Zhu, Y., et al. (2024a). *Unidexfpm: universal dexterous functional pre-grasp manipulation via diffusion policy*. arXiv preprint arXiv:2403.12421.

Wu, T., Wu, M., Zhang, J., Gan, Y., and Dong, H. (2023). Learning score-based grasping primitive for human-assisting dexterous grasping. *Adv. Neural Inf. Process. Syst.* 36, 22132–22150. Available online at: https://proceedings.neurips.cc/paper_files/paper/2023.

Wu, T., Wu, M., Zhang, J., Gan, Y., and Dong, H. (2024b). Learning score-based grasping primitive for human-assisting dexterous grasping. *Adv. Neural Inf. Process. Syst.* 36. Available online at: https://proceedings.neurips.cc/paper_files/paper/2024.

Xian, Z., Gkanatsios, N., Gervet, T., Ke, T.-W., and Fragkiadaki, K. (2023). "ChainedDiffuser: unifying trajectory diffusion and keypose prediction for robotic manipulation," in 7th Annual Conference on Robot Learning.

Xu, M., Xu, Z., Chi, C., Veloso, M., and Song, S. (2023). XSkill: Cross embodiment skill discovery. *Proc. 7th Conf. Robot Learn.* 229, 3536–3555. Available online at: https://proceedings.mlr.press/v229/xu23a.html.

Xu, Y., Mao, J., Du, Y., Lozáno-Pérez, T., Pack Kaebling, L., and Hsu, D. (2024). *"set it up!": functional object arrangement with compositional generative models*. arXiv preprint arXiv:2405.11928.

Yang, S., Du, Y., Kamyar, S., Ghasemipour, S., Tompson, J., Kaelbling, L., et al. (2024). "Learning interactive real-world simulators," in The Twelfth International Conference on Learning Representations.

Yang, Z., Mao, J., Du, Y., Wu, J., Tenenbaum, J. B., Lozano-Pérez, T., et al. (2023). Compositional diffusion-based continuous constraint solvers. *Proc. 7th Conf. Robot Learn.* 229, 3242–3265. Available online at: https://proceedings.mlr.press/v229/yang23d.html.

Ye, Y., Gupta, A., Kitani, K., and Tulsiani, S. (2024). "G-HOP: generative hand-object prior for interaction reconstruction and grasp synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1911–1920. doi:10.1109/cvpr52733.2024.00187

Yu, P., Xie, S., Ma, X., Jia, B., Pang, B., Gao, R., et al. (2022). Latent diffusion energy-based model for interpretable text modelling. *Proc. 39th Int. Conf. Mach. Learn.* 162, 25702–25720. Available online at: https://proceedings.mlr.press/v162/yu22h.html?ref=https://githubhelp.com.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., et al. (2020). Meta-world: a benchmark and evaluation for multi-task and Meta reinforcement learning. *Proc. Conf. Robot. Learn.* 100, 1094–1100. Available online at: https://proceedings.mlr.press/v100/yu20a

Yu, T., Xiao, T., Stone, A., Tompson, J., Brohan, A., Wang, S., et al. (2023). Scaling robot learning with semantic data augmentation through diffusion models. *Robotics Sci. Syst.* doi:10.48550/arXiv.2211.04604

Zare, M., Kebria, P. M., Khosravi, A., and Nahavandi, S. (2024). A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Trans. Cybern.* 54, 7173–7186. doi:10.1109/tcyb.2024.3395626

Ze, Y., Yan, G., Wu, Y.-H., Macaluso, A., Ge, Y., Ye, J., et al. (2023). GNFactor: multi-task real robot learning with generalizable neural feature fields. *Proc. 7th Conf. Robot Learn.* 229, 284–301. Available online at: https://proceedings.mlr.press/v229/ze23a.html.

Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H., et al. (2024). 3D diffusion policy: generalizable visuomotor policy learning via simple 3D representations. *Robotics Sci. Syst.* doi:10.48550/arXiv.2403.0395

Zeng, Y., Wu, M., Yang, L., Zhang, J., Ding, H., Cheng, H., et al. (2024). LVDiffusor: distilling functional rearrangement priors from large models into diffusor. *IEEE Robotics Automation Lett.* 9, 8258–8265. doi:10.1109/LRA.2024.3438036

Zhang, E., Lu, Y., Wang, W., and Zhang, A. (2024a). "Language control diffusion: efficiently scaling through space, time, and tasks," in International Conference on Learning Representations.

Zhang, F., and Gienger, M. (2025). *Affordance-based robot manipulation with flow matching*. arXiv preprint arXiv:2409.01083.

Zhang, J., Liu, H., Li, D., Yu, X., Geng, H., Ding, Y., et al. (2024b). "DexGraspNet 2.0: learning generative dexterous grasping in large-scale synthetic cluttered scenes," in 8th Annual Conference on Robot Learning.

Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., et al. (2024c). *NaVid: video-based VLM plans the next step for vision-and-language navigation*. arXiv preprint arXiv:2402.15852.

Zhang, J., Zhang, Y., An, L., Li, M., Zhang, H., Hu, Z., et al. (2024d). *ManiDext: hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion*. arXiv preprint arXiv:2409.09300.

Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., et al. (2024e). *VLABench: a large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks*. arXiv preprint arXiv:2412.18194.

Zhang, X., Chang, M., Kumar, P., and Gupta, S. (2024f). Diffusion meets DAgger: supercharging eye-in-hand imitation learning. *Robotics Sci. Syst.* doi:10.48550/arXiv.2402.17768

Zhang, Y., Gu, J., Wu, Z., Zhai, S., Susskind, J., and Jaitly, N. (2023). PLANNER: generating diversified paragraph via latent language diffusion model. *Adv. Neural Inf. Process. Syst.* 36, 80178–80190. Available online at: https://proceedings.neurips.cc/paper_files/paper/2023.

Zhang, Z., Wang, H., Yu, Z., Cheng, Y., Yao, A., and Chang, H. J. (2025). Nl2contact: natural language guided 3d hand-object contact modeling with diffusion model. *Comput. Vis. – ECCV* 2024, 284–300. doi:10.1007/978-3-031-73390-1_17

Zhang, Z., Zheng, K., Chen, Z., Jang, J., Li, Y., Wang, C., et al. (2024g). *GRAPE: generalizing robot policy via preference alignment*. arXiv preprint arXiv:2411.19309.

Zhang, Z., Zhou, L., Liu, C., Liu, Z., Yuan, C., Guo, S., et al. (2024h). *DexGrasp-Diffusion: diffusion-based unified functional grasp synthesis method for multi-dexterous robotic hands*. arXiv preprint arXiv:2407.09899.

Zhao, Y., Bogdanovic, M., Luo, C., Tohme, S., Darvish, K., Aspuru-Guzik, A., et al. (2025). *AnyPlace: learning generalized object placement for robot manipulation*. arXiv preprint arXiv:2502.04531.

Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., et al. (2024). 3D-VLA: a 3D vision-language-action generative world model. *Proc. 41st Int. Conf. Mach. Learn.* 235, 61229–61245. Available online at: https://dl.acm.org/doi/abs/10.5555/3692070.3694603.

Zhong, T., and Allen-Blanchette, C. (2025). *GAGrasp: geometric algebra diffusion for dexterous grasping*. arXiv preprint arXiv:2503.04123.

Zhou, H., Blessing, D., Li, G., Celik, O., Jia, X., Neumann, G., et al. (2024a). Variational distillation of diffusion policies into mixture of experts. in The thirty-eighth annual conference on neural information processing systems.

Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.-Y., and Gan, C. (2024b). RoboDreamer: learning compositional world models for robot imagination. in Forty-first international conference on machine learning.

Zhou, S., Du, Y., Zhang, S., Xu, M., Shen, Y., Xiao, W., et al. (2023). Adaptive online replanning with diffusion models. *Adv. Neural Inf. Process. Syst.* 36, 44000–44016. Available online at: https://proceedings.neurips.cc/paper_files/paper/2023.