Diffusion Model in Robotics: A Comprehensive Review

Abstract

Diffusion models are a powerful class of generative models that emerge in recent years to transform Gaussian noise into samples of the target distribution through an iterative denoising process. Due to the high training stability and powerful generative capabilities, diffusion models have surpassed previous generative models and demonstrate potential applications in the field of robotics. In the past few years, this area has gained increasing attention, and the number of studies applying diffusion models to the field of robotics has grown exponentially. This review aims to provide an overview of this emerging field, helping researchers understand the current state of development, with the hope of inspiring new research directions. First, we overview the foundation of diffusion models along with their development in recent years. On this basis, we provide an overview of the application of diffusion modeling in robotics from five aspects: scaling up robotics data, reinforcement learning(RL), imitation learning(IL), task planning and reasoning and other applications. We discuss and summarize the innovations, contributions, and limitations of these works. We then discuss the limitations and challenges faced by the field in terms of safety issues, real-time inference and model size, simulation to the real world gap, datasets and unified benchmarks and embodied foundation models. Finally, we summarize the review and provide an insight into future research directions.

Keywords: Diffusion Models, Robotics, Scaling up Robotics Data, Reinforcement Learning, Imitation Learning, Task Planning.

1. Introduction

As a new paradigm of artificial intelligence (AI), generative artificial intelligence (GAI) has received a great deal of attention in recent years. In contrast to traditional AI, GAI uses a large amount of data to train a generative model, which then outputs new data with a similar distribution to the training data. The generated data encompasses a diverse range of modalities, such

as text, image, and video. Due to the promising application prospects, generative models such as GPT[1], DALL-E[2], and Sora[3] have gained widespread attention. With remarkable success, GAI is applied extensively in industry, healthcare, education, and other domains. Therefore, researchers have also generated interest and ideas for applying GAI to the field of robotics. Despite Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs) [4], [5], [6], [7], [8] are attempted for the applications[4],[5],[6],[7],[8] in robotics, these models have not advanced due to their limitations in terms of generative capacity and training stability.

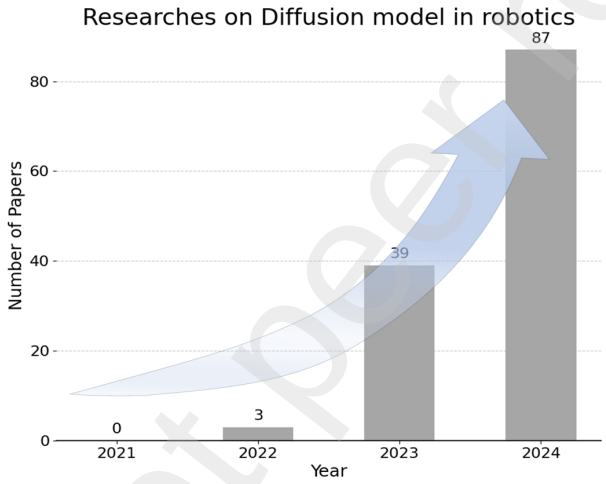

Figure 1: The number of researches applying diffusion models to robotics. Search for the keywords "Diffusion Models" + "Generative Models" + "Robotics" in ArXiv.

Diffusion models are a powerful class of generative models that transform Gaussian noise into samples of the target distribution through iterative denoising processes. These models have been widely used in signal processing, image generation, and image restoration due to their powerful generation ability and solid training stability. In order to capitalize on the potent generative powers of diffusion models and advance the field of robotics, an increasing amount of research in recent years has concentrated on integrating diffusion models with robotics, as shown in figure 1.

However, due to the diverse applications of diffusion models in the field of robotics, it remains challenging to fully understand how the models promote the development of robot learning in recent years. Therefore, a comprehensive review is conducted to fully investigate the practical applications of

diffusion models in robotics. While an existing review[9] has summarized the applications of diffusion models in planning and a recent work[10] has documented the utilization of generative models in robotics, our study distinctively provides a systematic examination of diffusion models across various aspects of robotics. This comprehensive review enables broader insights into the methodological implementations and practical potentials within the field. We hope this review will help researchers gain knowledge about the application of diffusion models in robotics, as well as the strengths and limitations, and possible future directions.

This paper is organized as follows: Section 2 describes the fundamental ideas of diffusion models and its latest developments. Section 3 provides a comprehensive overview of the applications of diffusion models in robotics by collecting and surveying research on the topic over the previous three years, including scaling up robotics data, RL, IL, task planning and reasoning and other applications. Section 4 analyses the limitations of applying diffusion models in robotics, addressing the challenges that are likely to arise in the future as well as potential research directions. Section 5 provides a summary of the key points discussed in this paper.

2. Overview of Diffusion Model

Diffusion models are a class of probabilistic generative models trained using variational inference to construct a Markov chain. Through iterative denoising processes, these models progressively transform Gaussian noise into samples from the target distribution, guided by the mean squared error (MSE) loss function. Owing to their robust capability to model complex distributions, diffusion models are widely applied in fields such as 2D image generation, video generation, and text generation. In this section, the denoising diffusion probabilistic model (DDPM)[11] is taken as a representative example to explain the fundamental principles of diffusion models. Next, we present a concise review of recent advancements in diffusion models, followed by a discussion of key techniques relevant to their applications in robotics.

2.1. Denoising Diffusion Probabilistic Models

As a landmark work, the publication of DDPM[11] has sparked an exponential growth of interest within the generative modeling community.

DDPM[11] is proposed based on previous works, with the fundamental concept of diffusion models first introduced in 2015[12]. Specifically, Diffusion

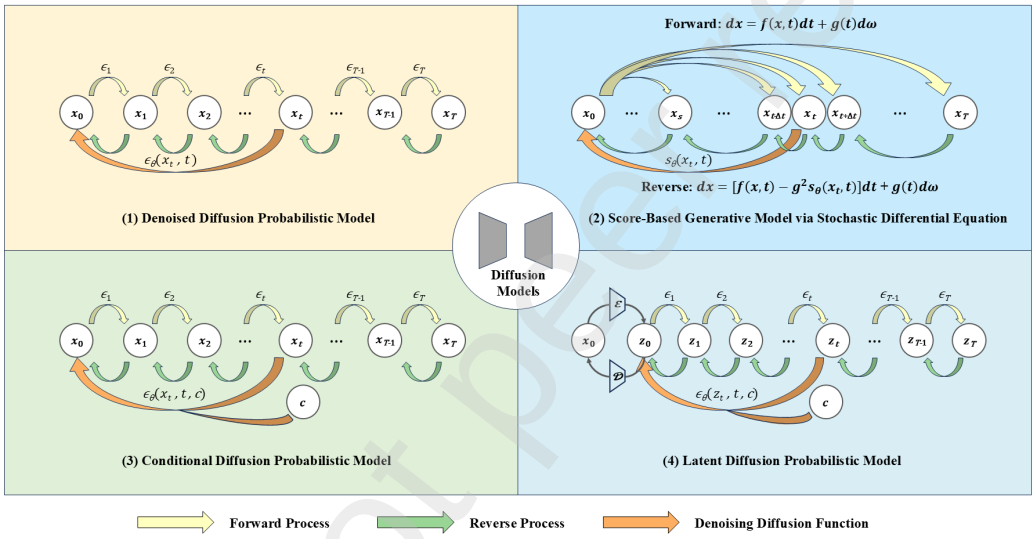

Figure 2: Overview of Diffusion Models. (1) The diffusion process in DDPM is implemented through iterative noise addition and removal. (2) Score-Based Generative Model governs the diffusion process through SDEs over continuous timeline (3) Conditional Diffusion Probabilistic Models employs condition c in each sampling step. (4)Latent Diffusion Probabilistic Model implements the diffusion process in a low-dimensional latent space.

Probabilistic Models[12] defines a forward process that diffuses a complex data distribution into a simpler one, and then learns the mapping between the two distributions by reversing this process.

The Score-based Generative Model[13], published in 2019, trains a shared neural network to estimate the score function, defined as the gradient of the log-density function of the perturbed data distribution. In this work, data is perturbed with Gaussian noise of varying magnitudes, and the score function is used to train the model, enabling the generation of denoised samples. These foundational works provide the theoretical basis for the development of DDPM[11].

Building upon the above works, diffusion model is further developed and optimized, and DDPM[11] is published in 2020. DDPM is defined as a parameterized Markov chain, where the target data distribution is generated through an iterative denoising process during inference time. DDPM is generally composed of two main components: the forward process and the reverse process.

Forward Process. In the forward process, DDPM is formulated as a Markov chain that gradually introduces Gaussian noise into the training data until the data is completely corrupted. Given data distribution $x_0 \sim q(x_0)$, where x0 refers to the original, uncorrupted training data, the noised versions $x_1, x_2, \ldots, x_T$ generated according to the following Markovian process:

$$q\left(x_{1:T} \mid x_0\right) := \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{2}$$

where T is the diffusion steps, and β_t is the hyperparameters representing the variance schedule across diffusion steps. An important feature of (1) (2) is that they enables direct sampling of x_t when t is drawn from a uniform distribution:

$$q\left(x_t \mid x_0\right) := \mathcal{N}\left(x_t; \sqrt{\hat{\beta}_t} x_0, \left(1-\hat{\beta}_t\right) I\right) \tag{3}$$

where $\hat{\beta}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$.

*Inverse Process.*With the above definition of the forward process, our objective is to generate new samples from noise that conform to the original data distribution. To achieve this, a reverse process can be trained using

learnable Gaussian kernels parameterized by θ:

$$p_\theta\left(x_{t-1} \mid x_t\right) := \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{4}$$

where μ_θ is the mean of learnable Gaussian kernels, and Σ_θ is the covariance. By iteratively denoising the data at time step t, it is possible to generate x_0 that follows the original data distribution.

The training objective of diffusion models is to minimize the variational lower bound (ELBO) of the negative log-likelihood, which has the following formulation:

$$\begin{aligned} \mathcal{L}_{vlb} =& KL\left(q\left(x_T \mid x_0\right) \| \pi\left(x_T\right)\right) - \log p_\theta\left(x_0 \mid x_1\right) \\ &+ \sum_{t>1} KL\left(q\left(x_{t-1} \mid x_t, x_0\right) \| p_\theta\left(x_{t-1} \mid x_t\right)\right) \end{aligned} \tag{5}$$

where KL denotes the Kullback-Leibler divergence between two probability distributions. In the total loss, the first term corresponds to the prior loss, which is independent of the parameters θ. The second term of the loss represents the reconstruction loss. The last term involves the true posterior of the forward process conditioned on the original data, and thus aims to minimize the deviation between the reverse process at each time step and the posterior of the forward process. It can be demonstrated that the posterior $q\left(x_{t-1} \mid x_t, x_0\right)$ follows a Gaussian distribution, which consequently leads to closed-form expressions for the KL divergences.

Furthermore, if we fix the covariance $\Sigma_\theta\left(x_t, t\right)$ to a constant value, we can express the mean $\mu_\theta\left(x_t, t\right)$ as a function of the noise, as follows:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \cdot \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\hat{\beta}_t}} \cdot \epsilon_\theta\left(x_t, t\right)\right). \tag{6}$$

Through the above simplification process, we can express $\mathcal{L}_{vlb}$ in equation (5) as follows:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t\sim[1,T]} \mathbb{E}_{x_0\sim p(x_0)} \mathbb{E}_{z_t\sim\mathcal{N}(0,\mathbf{I})} \left\|\epsilon_t - \epsilon_\theta\left(x_t, t\right)\right\|^2 \tag{7}$$

where $\mathbb{E}$ is the expected value, and $\epsilon_\theta\left(x_t, t\right)$ is the noise predicted by the network in x_t. This formulation measures the distance between the actual noise and the predicted noise at a random time step t. Through training,

the prediction network θ is employed in the reverse process for ancestral sampling.

Score SDE[14] further extends the discrete diffusion process in DDPM to the continuous case. In DDPM[11], the diffusion process gradually adds noise through a series of discrete steps, whereas in Score SDE[14], this process is modeled as a continuous stochastic process, which can be described using a stochastic differential equation. The continuous formulation allows for finer control over the diffusion process, potentially improving the quality and diversity of the generated samples.

Additionally, Song et al.[14] demonstrate that ordinary differential equations (ODEs) can also be employed to model the reverse process. Unlike SDEs, the probability flow ODEs, which are devoid of stochastic components, enable the use of larger step sizes during the integration process. The primary advantage of ODEs lies in their greater computational efficiency. For instance, methods such as PNDMs [15] and DPM-Solver [16] have achieved significantly faster sampling speeds by leveraging ODE solvers.

2.2. Conditional Diffusion Probabilistic Models

Although the generative process of DDPM is authentic, it is inherently random. To generate specific types of data using a diffusion model, it is necessary to train the model on a dataset of the desired type. However, this approach limits the flexibility of the model, thereby restricting its applicability. The introduction of conditional diffusion models enables the generation of arbitrary desired data, overcoming this limitation.

Conditional diffusion Probabilistic models can be expressed as $p_\theta(x|c)$, where c represents a given condition, such as a class label or text associated with the data x[17]. In the training of conditional diffusion models, there are two main sampling algorithms: classifier-free guidance[18] and classifier guidance[19].

Classifier Guidance. The core idea of the classifier-guided diffusion model is to adjust the noise prediction of the diffusion model during the diffusion process by using a pre-trained classifier, thereby increasing the likelihood of generating data from the target class. Specifically, at each diffusion step, the classifier computes the gradient of the probability that the sample belongs to the target class. The gradient of the output is then used to adjust the noise prediction of the diffusion model, making the sampling process more likely to generate data from the target class.

Classifier-free Guidance. In contrast to the classifier-guided diffusion model[19], the classifier-free guidance model eliminates the reliance on a separately trained classifier and instead uses the generative model itself to guide the sampling process. Specifically, classifier-free guidance jointly trains a single model with both the unconditional score estimator $\epsilon_\theta(x)$ and the conditional score estimator $\epsilon_\theta(x, c)$, where c denotes the class label. A null token $\varnothing$ is used as the class label in the unconditional part, $\epsilon_\theta(x, \varnothing)$. Experimental results in [18] demonstrate that classifier-free guidance strikes a balance between quality and diversity. Since the classifier-free guidance diffusion model does not require an additional classifier part, it is simpler and more flexible to implement, making it a widely used approach in conditional diffusion models.

The emergence of conditional diffusion models represents a significant milestone in the development of diffusion models. It further enhances the generative capacity and flexibility of diffusion models, driving their widespread application in various fields. This development has also attracted an increasing number of researchers to integrate diffusion models with robotics, thereby advancing the field of robotics.

2.3. The Recent Development

Although diffusion models have achieved significant progress in generative tasks, they still face a range of challenges, including high computational costs, slow generation speeds, sensitivity to hyperparameters, and the need for large datasets and considerable computational resources. In recent years, numerous studies have focused on addressing these challenges, aiming to improve the generative performance of diffusion models and expand their applicability to a wider variety of domains.

2.3.1. Diffusion Process

Traditional diffusion models treat the input noise at each step of the forward process as a wiener process in the pixel space[20]. However, this approach may be suboptimal for generative modeling. As a result, much research has focused on designing novel diffusion processes to simplify the reverse process and enhance the generative capabilities of diffusion models.

The latent diffusion models (LDMs) represent an improved version of the diffusion model, wherein the diffusion process occurs in a low-dimensional latent space. The frameworks of LDMs and various diffusion models are presented in figure 2. Classical works such as LSGM[21] and INDM[22]

jointly train a diffusion model alongside a VAE or a Normalizing Flow Model. The VAE or Normalizing Flow encoder is used to map images into a low-dimensional latent representation, which is then fed into a neural network. The decoder is subsequently used to reconstruct the output. Compared to diffusion models operating in pixel space, LDMs significantly reduce memory requirements and computational complexity, as the neural network is trained and inferred in the low-dimensional latent space. Stable Diffusion[17] is a notable example of latent diffusion models, in which a VAE is employed to map pixel-level representations to latent representations, and the diffusion model is conditioned on textual input during training.

Several works have focused on designing more robust and efficient forward processes. The Poisson Flow Generative Model[23] models ODE as high-dimensional electric field lines, with data points represented as charges in an augmented space. These charges move along the field lines, causing simple data distributions to gradually approach the target data distribution. Cold Diffusion[24] introduces image transformations, such as blurring and downsampling during the forward process, further enhancing the diversity of the images generated by diffusion models. Lipman et al.[25] employed a more advanced Gaussian noise kernel, making the training and sampling processes of diffusion models more efficient.

2.3.2. Sampling Acceleration

Generating data samples using diffusion models typically requires iterative methods, involving a large number of computational steps, which results in high time complexity. To address this issue, recent works have focused on accelerating the sampling process of diffusion models while ensuring that the quality of the generated samples is not significantly compromised.

Knowledge distillation[26] is a technique for transferring the knowledge from large, complex models to simpler models. In diffusion models, an ODE-based framework is typically employed to establish a direct mapping between the data distribution and the original noise distribution, enabling the student model to generate samples with fewer steps and a simpler network. Salimans et al.[27] firstly apply this method to accelerate sampling prosess by progressively distilling the sampling trajectory, where the teacher model iterates twice for every iteration of the student model, effectively halving the number of sampling steps. Meng et al.[28] proposed a two-stage distillation strategy to address the challenge of distilling knowledge from classifier-free guided conditional diffusion models. Further work[29][30][31] has focused on

directly estimating the target distribution from the noise samples at time step T, leading to the further increase in sampling efficiency. Unlike direct distillation methods, Reflow[32] is a trajectory distillation approach that enhances generation speed by improving the ODE of the teacher model. By minimizing the transport cost between distributions, it enables one-step generation.

Several works focus on improving ODE solvers or accelerating SDE sampling to enhance the efficiency of diffusion model. These methods use advanced samplers to accelerate the sampling process of pre-trained diffusion models, eliminating the need for retraining, which are collectively known as Training-Free Sampling. For instance, DPM-solver [16] extends DDIM[33] by solving a piecewise continuous ODE and reduces error accumulation using higher-order solvers, enabling the model to achieve results closer to the true distribution with fewer sampling steps. EDM[20] solves the ODE using a second-order Heun method, striking the balance between computational cost and accuracy. While SDE-based samplers are slower, they accumulate less error and provide better sampling quality. Gotta Go Fast[34] accelerates SDE sampling through adaptive step sizes. Meanwhile, the Restart Sampling algorithm[35] explores multiple generative paths using a restart mechanism, combining the advantages of ODE solvers, which have minimal discretization error, and SDE samplers, which have minimal accumulation error, while maintaining high sampling speed.

2.3.3. Likelihood Optimization

DDPM[11] optimizes the lower bound of the log-likelihood using ELBO, which helps to mitigate the challenges associated with directly optimizing the log-likelihood. However, likelihood optimization remains a difficult task. Therefore, some works have attempted to further achieve likelihood optimization by combining methods from Maximum Likelihood Estimation.

ScoreFlow[36] utilizes flow models to provide an initial approximate distribution, while the diffusion model corrects this approximation, jointly optimizing the log-likelihood of the data distribution. VDM[37] proposes a variational framework for diffusion models, aligning the time-continuous distribution in the diffusion process with the sampling distribution of the generative model. Huang et al.[38] reconstruct the training process of diffusion models from a variational perspective, interpreting the noise prediction task as an optimization problem within the variational framework. Additionally, some works[39][40] have achieved likelihood optimization by designing hybrid

loss functions.

2.3.4. Applications

Due to their powerful data modeling capabilities and realistic generative performance, diffusion models are now widely applied across various fields. These applications include, but are not limited to, image generation, video generation, 3D generation and text generation.

Image generation is the most mature application of diffusion models, and it can be categorized into unconditional generation and conditional generation. In unconditional generation, images are entirely controlled by the model's intrinsic distribution, resulting in randomness and diversity. In conditional generation, the process is guided by additional inputs, such as text descriptions, images, or other labels, enabling the generation of high-quality images that align with specific semantic information. Text-to-image diffusion models represent one of the most widely applied type of conditional diffusion models, with notable works including Imagen[41], Stable Diffusion[17], and DALL-E 2[2]. In the field of robotics, text-to-image diffusion models can generate visual representations based on textual prompts, assisting robots in task reasoning and planning. Furthermore, text-to-image models can generate a large volume of diverse training data, effectively alleviating the issue of data sparsity in the field of robot learning. Additionally, several works also focus on image generation under multimodal instructions and the application of video generation in robotics.

Given the ability to model and sample complex data distributions in a unified latent space, as well as their capacity for cross-modal generation, diffusion models are also capable of learning the relationships between images and action trajectories through joint modeling. In this context, diffusion models can be regarded as policies. In recent years, a substantial body of work has focused on the application of diffusion models in robot action sequences and motion trajectories generation, successfully sampling high-quality, temporally consistent targets that align with multimodal distributions. Currently, diffusion models have become an essential tool for addressing tasks in robot visual perception and motion planning.

3. Diffusion model in robotics

In this section, the application of diffusion models in various aspects of robot learning is discussed in detail, and the current state of development

and future directions are analyzed. First, We review the application of diffusion models to scaling up robotics data. Next, we elaborate on their distinct advantages in robot control from both reinforcement learning and imitation learning perspectives. Furthermore, we discuss how diffusion models enable novel capabilities in robotic task planning and reasoning. Finally, we provide a brief overview of other applications of diffusion models in the robotic domain.

3.1. Scaling up Robotics Data

The field of AI has been booming in recent years. As a key force leading the future development of science and technology, it is changing our lives at an unprecedented speed. The development of AI cannot be separated from three aspects, algorithms, arithmetic and data, which together drive the breakthroughs and innovations in AI technology. Countless scientific studies over the past decade have proven that large neural networks can produce remarkable results when combined with large datasets. Recent advancements in natural language processing (NLP) and computer vision (CV) have been primarily driven by the availability of large-scale, publicly accessible datasets collected from web sources.

In the field of robot learning, data sparsity is a key factor that limit progress. Many studies focus on algorithm innovation, demonstrating the robustness of algorithms only in limited environmental contexts, with a few objects to interact with and a small set of action skills. However, even state-of-the-art algorithms cannot achieve generalized embodied intelligence if a large amount of diverse data is missing. The desired data cover not only a wide range of motor skills, but also a wide range of different objects and visual domains. Several works[42][43] have demonstrated that using the same algorithms, robot learning algorithms perform more robustly and accurately when scaled to larger, more diverse datasets. When combined with large models of robot learning and large datasets, it is expected that higher performance robotic systems can be realized that can perceive, reason, and act in complex environments.

However, the acquisition of real-world robotics data is costly and time-consuming. The three usual methods for acquiring real-world robot data are kinesthetic teaching, teleoperation, and passive observation. All these methods require the involvement of experts and are cumbersome and time-costly. Moreover, Making the training data cover a wide range of objects and scenes requires extensive physical resources. To address these issues, past

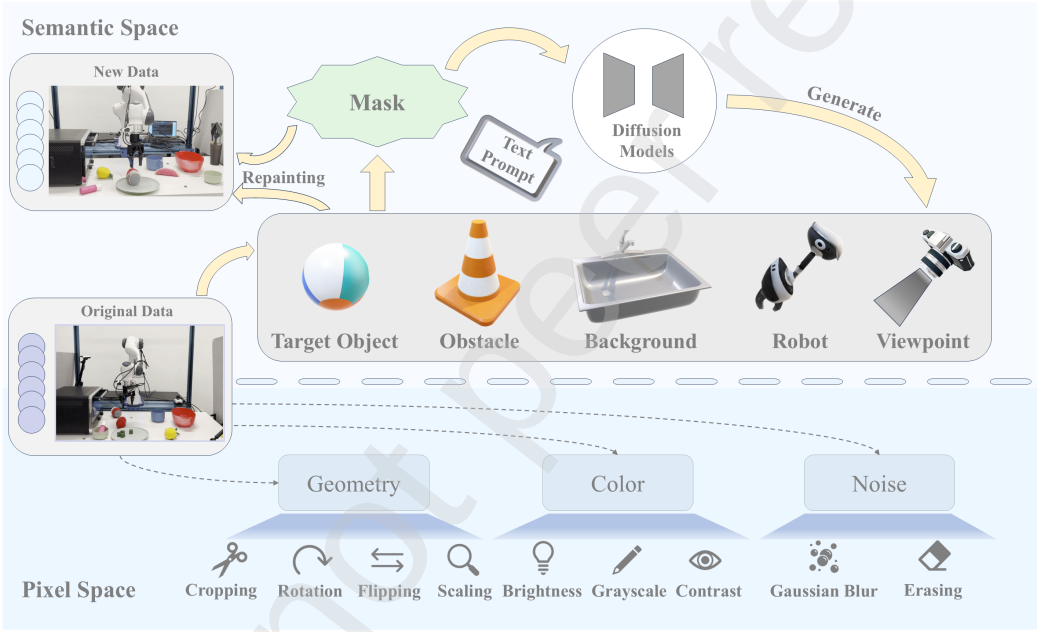

Figure 3: Data augmentation in robot learning. Traditional data augmentation methods perform simple pixel-level transformations in pixel space. However, diffusion models generate new valuable data in semantic space.

work has typically expanded the data volume using simulated data, simple data augmentation, and domain randomization. While building simulated environments makes it easy to introduce different objects, backgrounds, and tasks, simulated environments differ from real environments in appearance and various parameters. And algorithms trained on these data have limited ability to generalize to the real world. Simple data augmentation scales up the data by cropping, flipping, color transformations, etc., but it essentially introduces no additional semantic information, and merely makes low-level changes to the appearance of the scene, which is insufficient for scene generalization. Domain randomization[44][45], on the other hand, varies the physical and rendering parameters in the simulation environment, but it is similarly constrained by the gap between the simulation and the real world.

The emergence of diffusion model opens up new ideas for the expansion of robotic data due to its powerful and realistic generation capabilities. As is shown in figure 3, unlike traditional data augmentation methods, diffusion models achieve semantic-level data augmentation through novel asset generation.

CACTI [46] uses diffusion models for the first time to generate offline datasets for robot imitation learning. CACTI[46] replays expert or agent demonstrations multiple times, with each replay involving manual changes to the layout of some obstacles in the scene, ensuring the data is sufficiently diverse. Next, a manual mask is applied to the obstacles that need to be replaced. The mask guarantees the stability of generation, leaving portions of the scene outside the mask unaltered, and stable diffusion is used to generate new objects, along with a text prompt pertaining to the created object. Experiments have demonstrated that the strategy of CACTI [46] enhances the algorithm's task completion success rate, as well as its ability to generalize to different scenarios. However this work is only applied in a kitchen environment and generation is limited to task-independent objects. GenAug [47] instead generates obstacles, task objects, and backgrounds. The method still involves masking the objects to be replaced, providing a text prompt, and using a text-to-image diffusion model for generation. However, GenAug[47] addresses some details in the generation process, such as the fact that cross-category generation might significantly alter the shape and size of objects, which could overlap with the robot's original trajectory. Therefore, this work first renders an object of the same size from a mesh dataset, without rendering visual details, and then uses a depth-aware diffusion model to generate the complete object. After generation, it checks whether the bounding boxes

Selected works	Year	Target Problem	Contribution	Simulation Platform and Dataset
CACTI[46]	2023	IL; Scaling up Data	The first application of diffusion models for asset generation	CACTI -Sim-100; Realworld Kitchen
GenAug[47]	2023	IL; Scaling up Data	Refining generation details; Combined generation of backgrounds, objects and distractors	LVIS; Real-world
ROSIE[48]	2023	IL; Scaling up Data	Automating the generation process	RT-1
DMD[49]	2024	IL; Date Augmentation	Generating images with different object alignments and viewpoints	VIME; Realworld Grabbing and Pushing
RoVi-Aug[50]	2024	IL; Scaling up Data	Generating new robots and new viewpoints	Robosuite Simulator; Berkeley UR5
SYNTHER[54]	2024	RL; Scaling up Date	Combining experience replay with diffusion models to expand the experience data	D4RL; DMC; OpenAI Gym; V-D4RL
Gen2Sim[55]	2024	RL; Scaling up Date	Automating the Generation of Simulation Environments, Tasks, and Reward Functions	IsaacGym and Twin Environment in Reality
RoboGen[58]	2024	RL; Scaling up Date	Decomposing tasks and selecting the best learning method using a large language model	Genesis

Table 1: Diffusion models in Scaling up Robotics Data

of obstacles and task objects overlap and discards the overlapped ones.

ROSIE[48] uses a target detection model with a segmentation head to automatically perform object detection and segmentation. It further utilizes a large language model to refine and generate the text prompts needed for data augmentation. This work employs Imagen[41] to generate new objects and backgrounds and thus has automated the generation process. DMD[49] enhances algorithm robustness in in-distribution scenarios by employing diffusion models to generate images with slight variations in object arrangements and viewing angles within identical scenes. These synthesized images are subsequently utilized for imitation learning to train the policy. The proposed method demonstrates improved success rates in both object grasping and pushing tasks compared to conventional approaches. RoVi-Aug[50] introduces a new generation approach by creating different robots and viewpoints, thereby generalizing the algorithm to different robots and new camera perspectives. For robot generation, the work segments the robot in the image using the Segment Anything Model[51] and applies a mask. A fine-tuned stable diffusion model[17] is used to generate the robot image, and the E2FGVI[52] inpainting model fill the masked area with suitable images. For viewpoint generation, a 3D perception diffusion model, ZeroNVS[53], is used to generate images from different viewpoints.

Although the above work greatly scales up the robot learning data, there are still some problems in practical applications. First, whether using text-to-image or image-to-image diffusion models, they only generate objects, backgrounds, viewpoints, and so on, without generating new robot actions or trajectories. Essentially, they do not help the robot learn new skills. Second, since the generation process involves multiple tasks, such as segmentation,

generation, inpainting, and policy training, the cascading of multiple models may lead to error propagation, affecting the final result. Additionally, the generation speed is relatively slow. Lastly, because images are generated frame by frame, there is some loss of temporal consistency between frames, causing small visual differences between them. Future work may focus on solving these issues.

Diffusion models also have promising applications in data augmentation in simulation environments. SYNTHER[54] combines experience replay with diffusion models to flexibly expand the experience data collected by agents. In offline settings, SYNTHER[54] uses diffusion models to augment small offline datasets, and this additional synthetic data can effectively train more complex networks. In online reinforcement learning, the additional data allows agents to use much higher update-to-data ratios, leading to increased sample efficiency.

Some works combine diffusion models with large language models to automate the generation of simulation environments, tasks, and reward functions, significantly reducing the labor costs of training RL policies. Gen2Sim[55] first uses GPT-4[56] to obtain a list of object categories related to operational objects, then searches for corresponding 2D images, and maps these 2D images to 3D meshes with appropriate physical parameters to obtain different object assets. 2D image searching is partially done with stable diffusion[17]. The lifting from 2D images to 3D assets relies on Zero-1-to-3[57], a diffusion model for object view synthesis, which is accomplished by minimizing re-reprojection error and maximizing the likelihood of its image renderings. In addition to this, this work also used GPT4[56] to generate specific task and reward functions, successfully train RL polices, and validate the effectiveness of the algorithms deployed to the real world in a digital twin environment. RoboGen[58] uses a similar approach to Gen2Sim[55] to automate the generation of simulated 3D object assets, task descriptions and reward functions. At the same time, this work decomposes the task into a series of sub-tasks and automates strategy learning by selecting the best learning methods, such as RL, motion planning, and trajectory optimization, ultimately generating the required training objectives, thus achieving automated strategy learning. Robot task learning by RoboGen[58] encompasses a broad range of activities and contexts, such as legged mobility, deformable object handling, and rigid and articulated object manipulation. Despite producing a variety of RL data, these techniques are still produced in simulated settings, and there is still a Simulation-to-Reality (Sim2Real) gap with the real world.

3.2. Diffusion Model in Reinforcement Learning

Reinforcement learning (RL) is one of the most commonly used methods in robot learning, where an agent learns the optimal policy through interaction with the environment by maximizing the reward function. Although RL has been widely used in the field of robotics, there are still many problems. Foremost, the application of online RL is limited because RL requires a large amount of interaction data in order to learn effective strategies. However, in the real world, robot exploration and task attempts can be very expensive and time consuming. While offline RL alleviates the issue of low sample efficiency, it still suffers from insufficient data. In addition, rewards are often delayed or sparse in robotic tasks, which means the robot will not receive timely feedback during the learning process. In complex robot tasks, the dimensions of state space and action space are particularly large, which not only increases the difficulty of robot learning, but also suffers from high computational complexity and memory consumption. In recent years, a series of works have used diffusion models to improve reinforcement learning algorithms with notable success.

Diffuser[59] is the first work to apply diffusion models to trajectory optimization in RL. It generates planned trajectories directly from sampling and iterative denoising of the diffusion model, alleviating the dilemma of suboptimal trajectory generation by model-base methods. In contrast to past model-base methods, Diffuser[59] does not emphasize autoregressivity or Markov property, but rather iteratively denoises a set of state-action pairs to sample trajectories in the plan. The denoising result at each time step satisfies local temporal consistency, and through multiple steps of denoising, this temporal locality gradually extends to global consistency. To ensure the optimality of the sampled trajectories, Diffuser[59] was designed as a classifier-guided diffusion model to transform the RL problem into a conditional sampling problem. Experiments have shown that this approach has the benefits of long-horizon planning, temporal compositionality, variable trajectory length, and task compositionality, making it a milestone work in the integration of diffusion models and RL.

Diffusion-QL[60] further exploits the powerful modeling capabilities of diffusion models for complex data distributions to capture multimodal action distributions in offline reinforcement learning. It proposes to split the diffusion loss objective into two terms, a behavioral cloning term and a Q-learning term. So that the diffusion model samples the same distribution as the training set and generates high-value actions. SfBC[61] proposes to de-

couple policy learning to address the problem of limited policy expressivity. It decouples a strategy into an expressive generative behavior model and an action evaluation model. By doing so, this work transforms explicitly learning a target distribution into learning a behavior model. And the use of a diffusion model ensures high-fidelity behavioral modeling and improves the expressiveness of the strategy. DiffCPS[62] simplifies the constrained policy search problem using the action distribution of diffusion-based policies. It employs a primal-dual method with function approximation to solve the diffusion-based constrained policy search problem, thereby enhancing the expressiveness of the policy. SRDP[63] uses state reconstruction signals to guide the diffusion policy. By employing a state learning approach, it alleviates the distribution shift caused by out-of-domain states. EDGI[64] proposes an equivariant diffusion model by considering real-world spatial, temporal, and alignment symmetries, and integrates the model into a planning algorithm. This algorithm can also softly break symmetry under specific task requirements, improving sample efficiency and generalization ability. IDQL[65] generalizes implicit Q-learning into an actor-critic method, where the choice of convex asymmetric critic loss induces a behavior-regularized implicit actor. It replaces the previous implicit Gaussian policies with a diffusion model-based policy extraction algorithm, achieving better results.

The above works represent the strategy as a diffusion model sampling denoising process, which utilizes the powerful data modeling capability of the diffusion model to capture multimodal distributions and significantly enhances the expressive power of the strategy.

Decision Diffuser[66] uses diffusion model as a conditional generative model for sequential decision-making problems. Unlike Diffuser[59], it employs classifier-free guidance and low-temperature sampling, applying an inverse dynamics model to obtain decision actions. This work allows efficiently capturing the best decision-making actions in the dataset by conditioning the diffusion model generation with a high reward function. It is also possible to condition on different trajectory information and thus sample trajectories that satisfy specific constraints and are able to reflect different skills.

Energy-Guided Diffusion Sampling[67] introduces a non-standardized energy function to guide the generation process of diffusion models, successfully applying it to offline RL problems. Specifically, this paper proposes a new training objective called comparative energy prediction that compares the energies in a set of noise-perturbed data samples , and uses their soft energy labels as a supervision to learn intermediate guiding signals during the

diffusion process. This approach guarantees convergence to accurate intermediate guidance signals with infinite model capacity and data samples. It also achieve more accurate generation in complex robot control tasks.

LDCQ[68] combines LDMs with offline RL. This work addresses the problem of splicing suboptimal trajectories as well as extrapolation errors[69] in Q-learning. LDCQ[68] performs a two-stage training process to learn a low-level policy and a high-level latent diffusion prior. Subsequently, based on the latent diffusion prior, the algorithm performs batch-constrained Q-learning to mitigate the extrapolation error while improving the policy. This work shows that suboptimal demonstrations in offline RL datasets contain a rich, multimodal latent space. This latent space can be effectively captured using temporal abstraction and powerful conditional generative models, leading to strong performance when directly applied in a simple batch-constrained Q-learning framework.

Offline RL learns strategies from offline data and does not interact with the environment, so a major challenge is the difficulty of generalizing learned policies to unseen tasks. Offline meta-reinforcement[70][71][72] learning uses context encoders to learn task representations and trains the policy through temporal difference learning, effectively addressing the issue of poor task generalization. However, this traditional approach is not stable enough for strategy optimisation, so MetaDiffuser[73] pre-trained a context encoder to capture task-relevant information, which is then projected into a conditional diffusion model to generate task-compliant trajectories. MetaDiffuser[73] designs a dual bootstrap module during the sampling process of the diffusion model to enhance the kinetic consistency and high reward of the generated trajectories. This work proposes a new solution for offline meta-reinforcement learning using a diffusion model that allows the algorithm to better adapt to dynamically varying and reward-changing environments.

With the same aim of improving the generalization of the algorithm to unseen tasks, AdaptDiffuser[74] introduces a planning method with self-evolutionary capabilities. It generates rich, composite expert data for goal-conditioned tasks using reward gradients as guidance, and then fine-tunes the diffusion model with high-quality data selected by a discriminator. Experiments show that this method outperforms both seen and unseen tasks, and demonstrates stronger adaptability on unseen tasks.

Trajectory optimization based on diffusion models faces expensive iterative sampling costs. For example, the decision-making frequencies of Diffuser[59] and Decision Diffuser[66] are recorded as 1.3Hz and 0.4Hz, re-

Selected works	Year	Target Problem	Contribution	Simulation Platform and Dataset
Diffuser [59]	2022	Model-Based RL	Pioneering the application of diffusion models to trajectory optimization in reinforcement learning	D4RL; Block Stacking
Diffusion-QL[60]	2023	Offline RL	Integrating diffusion models with offline reinforcement learning, employing Q-value guidance to steer the diffusion learning	D4RL
SfBC[61]	2023	Offline RL	Decoupling a strategy into an expressive generative behavior model and an action evaluation model	D4RL
DiffCPS[62]	2023	Offline RL	Developing a primal-dual method to address constrained policy search in diffusion models	D4RL
SRDP[63]	2024	Offline RL	Employing state reconstruction signals to guide diffusion policies, mitigating distribution shift caused by out-of-distribution states	D4RL
EDGI[64]	2024	Model-Based RL	Introduceing an equivariant diffusion model that improves sample efficiency	3D Navigation; Block Stacking
IDQL[65]	2023	Implicit Q-Learning	Replacing implicit Gaussian policies with diffusion-based policies, strengthening Implicit Q-learning performance	D4RL
Decision Diffuser[66]	2023	Offline RL	Employs classifier-free guidance with low-temperature sampling, applying inverse dynamics models to derive decision actions	D4RL
Energy-Guided Diffusion Sampling[67]	2023	Offline RL	Proposing a new training objective called comparative energy prediction to guide the generation process of diffusion models	D4RL
LDCQ[68]	2024	Offline RL	Integrates latent diffusion models with offline reinforcement learning, reducing suboptimal trajectories and extrapolation errors	D4RL
MetaDiffuser[73]	2023	Offline Meta-RL	Introduces diffusion models as conditional planners to enhance Offline Meta-RL performance and generalization capability	2D Navigation; Multi-Task MuJoCo
AdaptDiffuser[74]	2023	Offline RL	Enhancing task generalization capability through adaptive synthetic data generation and diffusion model fine-tuning	D4RL; KUKA Robot
DiffuserLite[75]	2024	Offline RL	Refining trajectory generation through coarse-to-fine planning to reduce the search space	D4RL;FinRL; Robomimic
HDMI[77]	2023	Offline RL; Hierarchical	Deploying a cascade framework where the Goal Diffuser discovers sub-goals and then synthesizes corresponding action sequences.	D4RL
Hierarchical Diffuser[76]	2024	Offline RL; Hierarchical	Implementing hierarchical planning with a high-level Sparse Diffuser and a low-level trajectory generator.	D4RL
MADiff[78]	2023	Offline RL; Multi-Agent	Introducing diffusion models to multi-agent offline reinforcement learning tasks for the first time	MPE;MA Mujoco; SMAC;MATP
DOM2[79]	2023	Offline RL; Multi-Agent	Enhancing policy performance by expanding the dataset through replication of trajectories exceeding return thresholds	MPE; MA MuJoCo
DIPO[80]	2023	Online RL	Pioneering the integration of diffusion models into online reinforcement learning	MuJoCo
CPQL[81]	2024	Online RL	Introducing a consistency policy framework that establishes a one-step mapping from noise to target policy	D4RL; dm_control
PolyGRAD[82]	2024	Online RL	Proposing the first autoregressive world model framework capable of generating full trajectories	MuJoCo
DACER[83]	2024	Online RL	Formulating the reverse process of diffusion models as policy functions, using Gaussian mixture models to fit policy distributions	MuJoCo

Table 2: Diffusion models in Reinforcement Learning

spectively. The low decision frequency makes diffusion-based planning challenging for real-world applications and difficult to apply to long-horizon tasks. To address this, DiffuserLite[75] generates coarse-to-fine trajectories through a plan refinement process, starting with coarse planning that considers only distantly spaced states, and then progressively refining the details of execution between these states. This work reduces the length of the sequences generated by diffusion models and simplifies the complexity of the probability distribution. It also significantly reduces the planning search space, making it easier for the planner to find high-performance trajectories. In practical control, only the first action of each step is executed, so the presence of coarse estimations in temporal intervals distant from the current state does not lead to performance degradation.

One basic approach to solving complex problems is to break them down into a series of simpler problems. DiffuserLite[75] employs a hierarchical approach to optimize sampling efficiency, where the distinctive strength of this method manifests in enabling the execution of complex long-horizon tasks. Hierarchical Diffuser[76] combines the advantages of hierarchical planning and diffusion-based planning, dividing the planning of long-horizon tasks into high-level and low-level parts. The high-level planner, Sparse Diffuser, uses a skipping planning strategy to achieve a larger receptive field with lower computational cost. The low-level planner is responsible for specific action planning between each subgoal, further refining the results of high-level planning. This approach outperforms Diffuser[59] in both training and planning time, making it more suitable for long-horizon tasks. HDMI[77] also uses a hierarchical diffusion model to solve the long-horizon decision making problem. It adopts a cascaded framework, where a goal diffuser identifies subgoals, and a trajectory diffuser generates action sequences. The trajectory diffuser is based on a transformer-based model, which captures long-horizon dependencies between sub-goals more effectively than U-Net structures.

Compared to single-agent tasks, multi-agent scenarios face challenges such as interdependencies, complex collaboration, incomplete information, and more complex state and action spaces. MADiff[78] is the first to apply diffusion models to multi-agent offline reinforcement learning tasks. This approach uses an innovative attention-based diffusion model architecture, which computes attention across multiple latent layers in model of each agent, enabling effective information exchange and integration of global information from all agents. It can be used either as a decentralized strategy that performs simultaneous teammate modeling or as a centralized controller. In

experiments, this approach is applied to robot control tasks, where multiple agents control different joints of a robot to make it run as fast as possible. DOM2[79] incorporates a diffusion model into the policy network and employs conservative Q-values for policy improvement. This work increases the amount of data by replicating trajectories in the dataset that have return values above a threshold to improve policy performance. This approach enhances the learning efficiency while mitigating the propensity of policies to become trapped in local optima in multi-agent scenarios.

Online RL involves the agent learning through real-time interactions with the environment, continuously updating its strategy based on immediate feedback. Unlike offline learning, the agent in online RL uses the latest data at each moment, requiring a balance between exploration and exploitation. In the field of robotics, online reinforcement learning can help robots adapt to dynamically changing environments, but it also faces high cost and safety issues. Recently, some studies have shown the potential of applying diffusion models to online RL. DIPO[80] is the first to apply diffusion models to this field. It proposes an action relabeling method that updates actions along the gradient field of the state-action value function. This action gradient optimization approach effectively improves the reward performance of the policy. CPQL[81] proposes a consistency strategy to realize a one-step mapping from noise to the desired policy, which significantly improves the inference speed. This approach can be seamlessly extended to online RL tasks without additional exploration. PolyGRAD[82] introduces the first non-autoregressive world model, capable of generating complete trajectories that align with the policy. It demonstrates excellent performance in short-horizon trajectory optimization. Traditional RL algorithms typically parameterize policies as diagonal Gaussian distributions. This representation constrains the expressive capacity, failing to capture the strong multimodal characteristics that theoretically optimal policies may exhibit. To this end, DACER[83] used the inverse process of the diffusion model as a policy function and proposed the use of a Gaussian mixture model to fit the strategy distribution. The entropy of the policy is estimated based on the GMM, and then a new parameter is learned, which is fine-tuned to regulate the degree of exploration and exploitation of the strategies.

Compared to offline reinforcement learning, there are fewer studies combining diffusion models with online reinforcement learning. While the works mentioned above are important explorations, they mostly focus on simple experiments in the MuJoCo[84] benchmark dataset. There is a lack of ex-

perimentation in real robot environments, which calls for further research.

3.3. Diffusion Model in Imitation Learning

Imitation learning (IL) is a learning method based on the existing demonstration. The agent learns knowledge from expert demonstrations and performs the same actions in the same environment. Compared to RL, IL does not require interaction with the environment or manually set reward functions, making it more efficient. Recently, many works have applied diffusion model combined with IL to the field of robot control. These works learn skills from a small number of samples and have made remarkable progress.

One of the major advantages of diffusion models in IL is their ability to capture multimodal action distributions. Some works [85][86] have demonstrated this advantage. With the conditioning mechanism, the models can flexibly select and combine distinct behaviors, and even generate policies for specific behaviors through descriptions. Representative works of diffusion models in imitation learning and trajectory generation are shown in Figure 4. Pearce et al.[87] are the first to apply diffusion models to behavior cloning for imitating human behavior in sequential environments. This work demonstrates the ability of diffusion models to model complex action distributions in robot control tasks and game environments. However, it also highlights issues such as increased inference time and numerous hyperparameters. Meanwhile, BESO[88] uses score diffusion models to generate actions. It decouples the score model from the inference sampling process, allowing for faster sampling of policies and significantly reducing the number of denoising steps. ChainedDiffuser[89] is another work that learns policy from demonstrations. This work combines high-level macroscopic action prediction with low-level trajectory diffusion generation, achieving good results even with limited demonstration data and monocular views. Diffusion Co-policy[90] uses a Transformer-based diffusion model to predict joint action sequences of humans and robots in collaborative tasks. By conditioning on historical human actions, the robot can better adapt to human behavior while also adjusting to changing environments.

Diffusion policy[91] represents the robotic visuomotor policy as a conditional denoising diffusion process to generate behaviors, drawing widespread attention in the robotics field. This work introduces a Transformer-based diffusion network that mitigates the over-smoothing effects commonly observed in Convolutional Neural Network (CNN). It utilizes backward horizon control and demonstrates superior performance in tasks that demand high-frequency

Selected works	Year	Target Problem	Contribution	Simulation Platform and Dataset
Pearce et al.[87]	2023	BC	The first application of diffusion models to BC, demonstrates effectiveness in imitating human behaviors.	D4RL;Kitchen environment
BESO[88]	2023	IL	The first work to utilize score-based diffusion models for action generation	CALVIN;Block-Push; Relay Kitchen
ChainedDiffuser[89]	2023	IL; Hierarchical	Using diffusion models to hierarchical framework for imitation learning to generate trojectories	RLBench; Real-world
Diffusion Co-policy[90]	2023	IL	Proposing a transformer-based diffusion model to predict joint action sequences in human-robot collaborative tasks	Motion Planning in Simulation and Real-world
Diffusion policy[91]	2023	IL	Formulating robotic visuomotor policies as conditional denoising diffusion processes to generate robot actions	Robomimic;Push-T; Multimodal Block Pushing; Franka Kitchen;Real-world
Consistency Policy[92]	2024	IL	Introducing consistency distillation that transfers teacher model knowledge to student policy.	Robomimic;Push-T; Franka Kitchen;Real-world
VDD[93]	2024	IL	Compressing diffusion model into mixture-of-experts framework, accelerating inference speed	Relay Kitchen;Block-Push; D3IL
SDP[94]	2024	IL	Proposing an action buffering mechanism that implements partial denoising trajectories, accelerating inference speed	Robomimic;Push-T; Real-world Push-T
BRIDGER[95]	2024	IL	Optimizing reverse process initialization that replaces Gaussian noise with source policy	Franka Kitchen;Adroit; 6-DoF Grasp;Real-world
Equivariant Diffusion Policy [96]	2024	IL	Incorporating equivariance into diffusion policy design, utilizing domain symmetry to boost training sample efficiency	MimicGen; Real-world
MDT[97]	2023	IL; Foundation	Combining multimodal Transformers with pretrained foundation models to learn long-horizon manipulation tasks	CALVIN
Crossway Diffusion[98]	2024	IL	Incorporating a self-supervised auxiliary objective alongside the standard diffusion loss to optimize the model	Robomimic; Push-T;Real-world
HDP[99]	2024	IL; Hierarchical	Integrating diffusion policy into hierarchical framework, leveraging differentiable dynamics to optimize action generation	RLBench; Real-world
DISCO[100]	2024	IL; Hierarchical	Using VLMs to generate keyframes which serve as coarse guidance in diffusion model for robotic motion generation	CALVIN;Block-Push; Franka Kitchen;Push-T; Grasp Pose;Real-world
Track2Act[101]	2024	IL; Hierarchical	Predicting plausible high-level plans using Internet videos and enable efficient fine-tuning with minimal robot data	Real-world
Xskill[102]	2023	IL	Proposing a skill-conditioned policy to acquire skills from unlabeled human demonstrations	Franka Kitchen; Realworld Kitchen
DP3[103]	2024	IL; 3D	Presenting the first work integrating 3D visual representations into diffusion policy	MuJoCo;Sapien;IsaacGym; PlasticineLab;Real-world
3D Diffuser Actor[104]	2024	IL; 3D	Encoding and concatenating 3D scene representations with text instructions and robot proprioceptive states	RLBench; CALVIN;Real-world
DNAct[105]	2024	IL; 3D	Using NeRF to pretrain 3D encoders, distilling 2D semantic features into 3D space via neural rendering	RLBench; Real-world
R&D[106]	2024	IL	Unifying robotic motions and RGB observations in a shared image space, making action align through diffusion models	RLBench; Real-world
NoMaD[107]	2024	IL; Navigation	Generating action sequences for navigation tasks using diffusion policy, achieving exploration and navigation mode switching	6 distinct indoor and outdoor environments
Yoneda et al. [108]	2023	IL	Ensuring fidelity of human inputs and congruence maintenance of target behaviors in human-robot collaborative control	D4RL;Open AI Gym; Lunar Reacher;Block Pushing;
RDT[109]	2024	IL; Foundation	Proposing the first diffusion foundation model for dual-arm robotic manipulation that overcomes data heterogeneity problem	46 Datasets of Various Robots for Pretraining;7 tasks for test
This&That [135]	2024	Task Planning; Text-Video	Introducing additional gesture cues on initial frames to deliver targeted guidance for diffusion model	Bridge;IsaacGym
SkillDiffuser [136]	2024	Task Planning; Text-Video	Encoding skills into finite skill sets via Vector Quantization and utilizing skill abstractions as conditioning inputs for diffusion models.	LOReL Sawyer; Meta-World
Botteghi et al. [142]	2023	Motion Planning; Safety	Trainning a DDPM with CBF constraints for trajectory safety classification and security evaluation	Openai Gym
SafeDiffuser [143]	2025	Motion Planning; Safety	Integrating CBFs into conditional generation process of diffusion model to enforce generated trajectory safety	D4RL
Lee et al.[144]	2023	Motion Planning; Safety	Propose a new recovery gap metric to evaluate the feasibility of plans generated by diffusion models.	Maze2D;Gym MuJoCo; Block Stacking
LTLDoG[145]	2024	Motion Planning; Safety	Incorporating LTLf into the diffusion model to ensure generated trajectories satisfy specified static and temporal constraints	Maze2D;Push-T; Real-world

Table 3: Diffusion models in Imitation Learning

action changes and precise speed control. These innovations allow diffusion policies to model multimodal action distributions, predict future action sequences, and maintain high training stability. However, Diffusion Policy also has some drawbacks, including its dependence on the quality of demonstration data, as well as higher computational costs and inference latency. Following works have extensively addressed and improved these challenges.

Consistency Policy[92] aims to further improve the inference speed of Diffusion Policy[91], enabling it to perform robot control tasks even in resource-constrained and dynamic environments. This work uses consistency distillation to transfer knowledge from a teacher model to a student model, enabling faster action generation in just one or three inference steps. VDD[93] also uses a distillation approach to distill the diffusion model into a mixture of experts. VDD[93] optimizes each expert by pre-training the diffusion model with a score function, which enables knowledge transfer between models through variational inference. This approach combines the strong generative capabilities of diffusion models with the efficient inference and interpretability of mixture of experts. Although the distillation technique reduces the number of steps required for action prediction, this approach still compromises sample quality and diversity. Instead, SDP[94] introduces an action buffer that generates partially denoised action trajectories, where the immediate action is noise-free, and the noise level gradually increases for subsequent actions. By recursively updating this buffer, SDP enables recursive sampling. This approach improves inference speed while keeping performance on par with traditional Diffusion Policy[91].

BRIDGER[95] accelerates inference speed by optimizing the initialization of the reverse diffusion process, resembling a warm-start approach. It creates a connection between the source policy and the target policy through random interpolation. The source policy, which can be a task-based heuristic or a data-driven policy, serves as the starting point of the diffusion process instead of the conventional Gaussian noise initialization.

Equivariant Diffusion Policy[96] introduces the concept of equivariance into diffusion policies, using domain symmetry to improve training sample efficiency and the generalization ability of diffusion models. This work theoretically proves that denoising diffusion functions are equivariant under certain conditions and demonstrates the application of SO(2)-equivariance in 6-Degree-of-Freedom (6-DoF) control for robotic manipulation. The authors then propose a new architecture that leverages equivariant networks to train the diffusion model and generates target action sequences through decoding

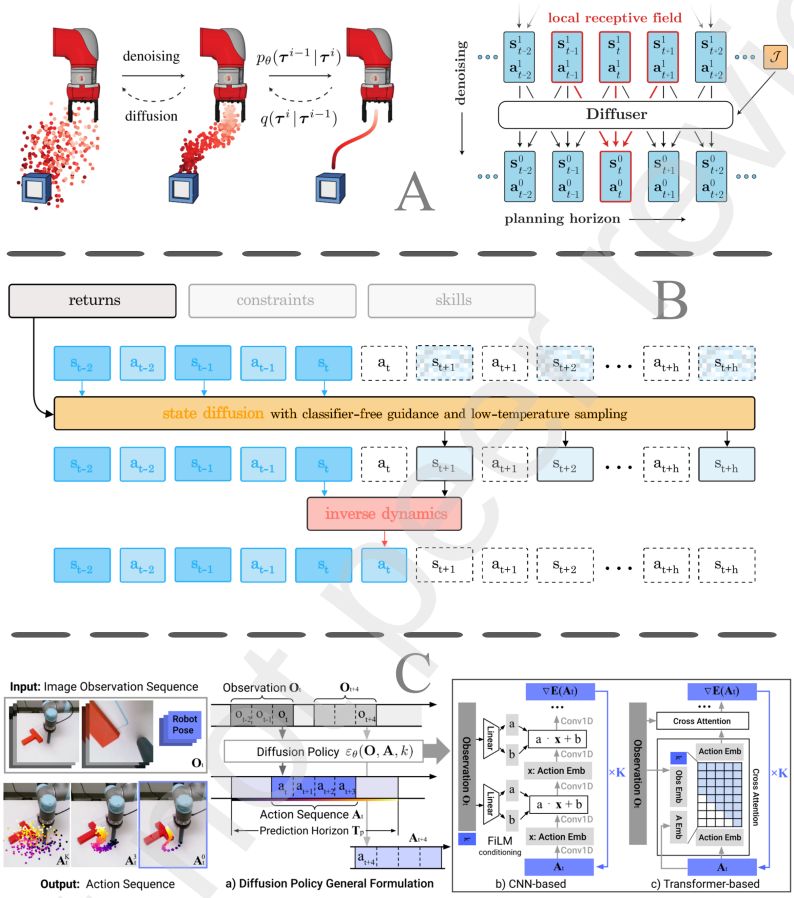

Figure 4: Diffusion models used in trajectory generation and policy learning. A. Diffuser samples plans by iteratively denoising two-dimensional arrays consisting of a variable number of state-action pairs (Diffuser[59]). B. Given the current state and conditioning, Decision Diffuser uses classifier-free guidance with low-temperature sampling to generate a sequence of future states. It then uses inverse dynamics to extract and execute the action (Decision Diffuser[66]). C. Diffusion Policy takes the latest steps of observation data as input and outputs several steps of actions. The backbone can be CNN or Transformer (Diffusion Policy[91]).

with equivariant embeddings. Equivariant Diffusion Policy achieves efficient action generation with a small amount of training data, significantly improving training sample efficiency.

Boosting inference speed allows diffusion policies to meet real-time requirements. While enhanced reasoning capabilities further empower these policies to flexibly handle action generation for complex tasks. Researchers improve performance of diffusion policy through three key approaches: introducing additional loss functions, adopting hierarchical architectures, and utilizing model inputs with stronger expressive capacity.

Crossway Diffusion[97] enhances model optimization by incorporating a self-supervised auxiliary objective alongside the standard diffusion loss. Specifically, the method introduces a state decoder that reconstructs raw image pixels and other states from intermediate representations during training. This joint supervision mechanism provides a straightforward yet effective solution for challenging tasks.

HDP[98] integrates diffusion policy with a hierarchical strategy. The policy consists of a high-level next-best-pose agent and a low-level goal-conditioned diffusion policy. Additionally, in the low-level policy, this work applies differentiable dynamics to distill the end-effector pose trajectory into a joint position trajectory, ensuring compliance with kinematic constraints. Simulations and real-world experiments verify that HDP generalizes better to long horizon tasks. It also effectively avoids errors from inverse kinematics solvers. The generated motions are kinematically accurate and feasible. DISCO[99] uses keyframes generated by visual language models (VLMs) as high-level guidance in the diffusion process. This method effectively generalizes to unseen tasks and tasks with open-vocabulary descriptions. However, the keyframe inference does not account for complex scenarios which encounter occlusion challenges. Track2Act[100] aims to train a model for general robot manipulation using internet data. It leverages online videos to predict high-level plans and uses a conditional diffusion process to forecast the future positions of key points. Given the depth image of an initial scene, the model can convert 2D images into 3D end-effector poses. With minimal fine-tuning on robot-specific data, the model can generalize to unseen tasks and scenes.

XSkill[101] modifies the goal-conditioned policy to a skill-conditioned policy. It proposes learning reusable skill prototypes from unlabeled human operation videos to create a shared skill representation space for both humans and robots. By using the skill prototype representation as a condition, the

diffusion policy is guided to generate action sequences. This approach enables skill transfer between different embodiments and allows learning new robot skills from a single human demonstration video. However, the ability to generalize to different backgrounds and viewpoints is still unknown.

In visuomotor control algorithms, 3D representations provide more expressive power than 2D ones, offering richer environmental information. DP3[102] is the first to apply 3D visual representations to diffusion policy. This approach uses monocular camera to acquire point cloud data, which is encoded into compact 3D representations through a lightweight Multilayer Perceptron (MLP) encoder, and then generates action sequences using a conditional diffusion denoising model. Comparisons with other 3D representations and encoders show that sparse point cloud is the most effective. DP3[102] can handle complex, high-dimensional tasks with training from only a few demonstrations. 3D Diffuser Actor[103] encodes and combines 3D scene representations, language instructions, and robot proprioception information as conditions to guide the diffusion model in action generation. However, the inference speed is slow due to the introduction of too much information. DNAct[104] pre-trains a 3D encoder using Neural Radiance Fields. DNAct then distills 2D semantic features from the foundation model into 3D space using neural rendering. This general 3D representation, which includes common-sense priors, is used as the condition for diffusion policy. This work is parameter-efficient and demonstrates strong generalization in both simulation and real-world environments.

The core challenge in visuomotor control for robots is the complex mapping between high-dimensional observations and low-level actions, especially when data volume is limited. R&D[105] utilizes a 3D model of the robot to unify the low-level robot actions and the RGB observations into a single image space. The rendered action representations are then iteratively updated by learning a denoising process until they are tightly aligned with the actions in the training data. This method simplifies the learning problem, enhancing both sample efficiency and spatial generalization. However, it comes with the trade-off of higher computational costs and a strong dependence on camera parameters and calibration.

In addition to single-arm robot control, diffusion models have also been widely applied to other robotic control tasks. NoMaD[106] focuses on robot navigation strategies, using diffusion policy to generate action sequences for navigation tasks. This work introduces a target mask mechanism that allows flexible switching between target and non-target scenarios. Through this

mechanism, it can handle both task-agnostic exploration and task-oriented navigation. Yoneda et al[107] implemented collaborative human-robot control in a robotic system based on the diffusion model. This work allows a trade-off between the fidelity of user actions and the consistency of target behavior by defining a forward diffusion ratio.

RDT[108] is a foundation model for dual-arm robot manipulation, using Diffusion Transformer as its backbone. The Transformer architecture enables the encoding of different modality inputs, eliminating the heterogeneity between visual inputs, language instructions, and other low-dimensional inputs. To build a universal foundation model across different robots, this work proposes a physically interpretable unified action space, which encodes various robots with different action spaces into this unified space. The unified space preserves the physical meaning of the original actions. RDT[108] is pretrained on 46 datasets and over one million trajectories, and then fine-tuned on more than 6000 dual-arm operation data collected in house-environment , significantly improving generalization ability.

3.4. Task Planning and Reasoning

In addition to generating actions, diffusion models exhibit strong reasoning capabilities when trained on sufficient data. They can assist robots in task planning by understanding and breaking down complex tasks. By leveraging these capabilities, robots can achieve more adaptive and intelligent behaviors in dynamic environments.

Carvalho et al.[109] are the first to propose conditioned score-based models to robotic motion planning tasks in complex environments. This work utilizes signed distance functions to encode the geometric properties of environmental maps, demonstrates promising performance in 2D navigation tasks, and lays the foundation for applications in robots with higher degrees of freedom. SE(3)-DiF[110] learns a smooth cost function in SE(3) space and trains the diffusion model through joint gradient optimization, achieving outstanding performance in 6-DoF grasping tasks. MPD[111] also utilizes a diffusion model to fit trajectory distributions and demonstrates its effectiveness in 3D maze navigation and 7-DoF robotic manipulation tasks. DiMSam[112] employs diffusion models to generate action sequences that adhere to constraints along with parameter values for each action, partially addressing the challenge of multi-step robotic operations in partially observable environments. Power et al.[113] utilize the trajectory distribution generated by diffusion

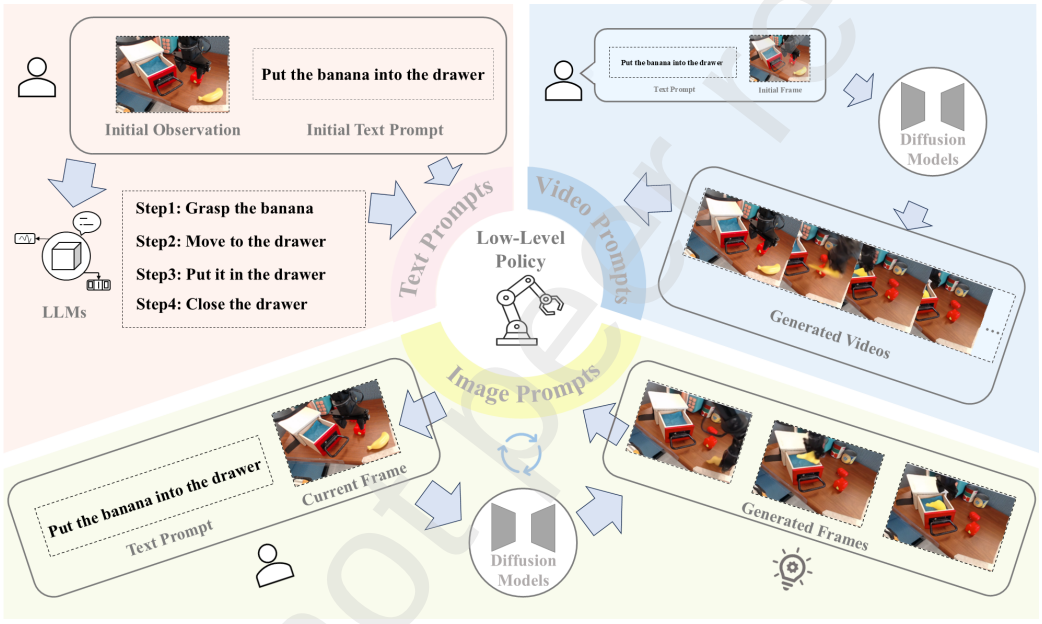

Figure 5: Three task reasoning methods. LLMs decompose tasks and output subgoal text prompts. While diffusion models generate video prompts or iteratively take current observations and text as input to generate a series of intermediate keyframes. Images and video possess far greater representational capacity than text, thus significantly enhancing task comprehension of agent.

models as an initialization for Constrained Stein Variational Trajectory Optimization, demonstrating generalization to unseen constraint combinations during training in a 12-DoF quadrotor task. EDMP[114] also uses a diffusion model to learn motion trajectory patterns. Their method proposes an Ensemble-of-Costs Guidance technique. This guidance helps generate trajectories that meet specific needs by combining different cost functions. Luo et al.[115] propose Potential-Based diffusion motion planning, which learns and combines potential fields to achieve generalization across diverse motion constraints.

In order to enable robots to autonomously execute complex high-level tasks, task planning is indispensable. In recent years, the application of diffusion models in task planning has also garnered significant attention. The implementation of diffusion models for task planning first emerges in the domain of object reorientation, where StructDiffusion[116] employes robots to accomplish object rearrangement tasks. This approach inputs point clouds and language instructions into diffusion models, enabling the rearrangement of unseen objects into physically plausible structures, thereby guiding the motion planning. DALL-E-Bot[117] employs a web-scale trained diffusion model for task planning, generating target images through natural language descriptions, which then guide robots to rearrange objects via grasping and placing operations. However, this work is only applied to 3-DoF rearrangement tasks. Subsequent research extends it to higher DoF object rearrangement tasks. ReorientDiff [118] employs a diffusion model to sample intermediate poses from a joint representation, enabling robots to achieve more precise object placement in both position and orientation. These existing studies primarily focus on object reorientation under explicit instructions, which imposes high demands on command accuracy and specification details. SetItUp[119] introduces a novel approach to interpret and execute ambiguous instructions. This framework utilizes large language models (LLMs) to first generate abstract spatial relationships between objects as geometric constraints. Subsequently, it combines multiple diffusion models to generate object poses that satisfy these constraints. In addition, RPDiff[120] approaches rearrangement prediction as a point cloud pose denoising problem. This method learns geometric relationships between objects and scenes through point cloud inputs. It generates multiple potential rearrangement configurations. However, this approach has limitations in requiring substantial training data. Obtaining 3D point clouds from real-world scenarios proves particularly challenging, which restricts its practical application in complex environments.

Most robotic manipulation tasks require satisfying multiple constraints. These constraints typically include robot limitations such as joint movement ranges, and environmental restrictions such as collision avoidance. To address this challenge, Diffusion-CCSP[121] proposes a constraint satisfaction framework for robotic reasoning and planning. This method employs factor graphs to visually represent different constraints. Specifically, a diffusion model is trained for each constraint type and transformed into an energy function. The energy functions collectively form a global optimization objective. Through iterative energy minimization sampling, solutions satisfying all constraints can be systematically generated. GSC[122] approaches robot task planning by constructing a skill chain. This method trains an unconditional diffusion model for each individual skill. During inference time, these models are sequentially combined to address unseen long-term goals. However, this approach has only been experimentally validated in fully observable environments, with no consideration given to partially observable scenarios. In contrast, Planning as In-Painting[123] frames task planning as an image in-painting problem. This method simultaneously estimates target states and trajectories through joint optimization. Such formulation effectively addresses partial observation challenges. The proposed framework enhances algorithmic adaptability in complex environments by implementing dynamic plan updates during execution. Diffusion-EDFs[124] proposes a bi-equivariant diffusion model based on the SE(3) space for task planning. This approach utilizes the invariance properties of translation and rotation. Such properties allow the model to be trained with limited data while maintaining generalization capabilities to unseen task configurations during testing. However, this study does not achieve control-level or trajectory-level reasoning. Equivariant Diffusion Policy[79] later addresses this limitation.

While diffusion models have been widely applied in generating action sequences and parameterized motions, their primary strength still lies in image generation. Recent studies demonstrate that image diffusion models can effectively synthesize goal states for robotic tasks. More importantly, these models enable the decomposition of complex tasks into sequential modular operations. As is shown in figure 5, the emergence of diffusion models has made visual prompts become a key tool in robotic task reasoning. Diverging from text and visual prompts, image prompts condition on camera-captured current frames to continuously generate subsequent frames upon completing the underlying control task. This closed-loop generative process enables dynamic task planning adjustments through iterative refinement cycles. Such

hierarchical approach has shown promising results in robotic manipulation scenarios, particularly in solving long-horizon tasks through stepwise visual planning. Experimental validations across multiple robotics platforms have confirmed the practical feasibility of this methodology for task decomposition and multi-stage execution.

SuSIE[125] fine-tunes the InstructPix2Pix[126] model to generate hypothetical future frames based on current visual frames and language descriptions. These frames serve as high-level guidance for task execution. A low-level controller then translates this guidance into concrete robot actions. By leveraging web-scale training data, the system demonstrates enhanced capabilities in recognizing and reasoning about novel objects and scenarios beyond those encountered in domain-specific datasets. The separation of high-level visual reasoning and low-level control enables effective generalization to unseen environments. GENIMA[127] introduces additional information on generated images to enhance system performance. This study fine-tunes the Stable Diffusion model[17] to predict future joint position targets based on input RGB images and linguistic objectives. The predicted joint positions are visualized as colored spheres superimposed on the images. These spheres contain horizontal stripes that explicitly indicate joint rotation angles through their spatial patterns. Experimental results demonstrate that this information fusion approach achieves superior performance compared to single-modality guidance methods. CoTDiffusion[128] employs a hierarchical control approach for robotic manipulation, using diffusion models as high-level visual planners. This method features a semantic alignment module that iteratively generates sub-goal images for subsequent steps. These images are aligned with multimodal instructions to provide clear semantic guidance for the diffusion process. Through rigorous semantic alignment, CoTDiffusion produces coherent sub-goal image sequences that closely follow task requirements.

UniPi[129] proposes a novel approach by transforming sequential decision-making problems into text-conditioned video generation tasks. This method employs pretrained language embeddings and leverages internet video datasets to synthesize highly realistic video plans. These generated video sequences serve as visual guidance for robot action execution. High-level planning based on videos provides more detailed guidance. However, it also has drawbacks, such as slow inference speed and the potential for generating unrealistic objects or motion hallucinations in partially observed environments. AVDC[130] employs a video generation approach to guide task execution.

Selected works	Year	Target Problem	Contribution	Simulation Platform and Dataset
Carvalho et al. [110]	2022	Motion Planning	The first work to propose conditioned score-based models to robotic motion planning tasks in complex environment	RECTANGLES AND CIRCLE
SE(3)-DiF [111]	2023	Motion Planning	Proposing a unified optimization framework for learning smooth cost functions to generate motion trajectories in SE(3) space	IsaacGym; Real-world
MPD[112]	2023	Motion Planning	Learning prior trajectory distributions using diffusion models	PointMass2D Dense; PointMass3D Maze Boxes
DiMSam[113]	2024	Motion Planning	Generating constraint-satisfying action sequences and corresponding parameters for each action using diffusion model	IsaacGym; Real-world
Power et al. [114]	2023	Motion Planning	Initializing the Constrained Stein Variational Trajectory Optimization with trajectory distributions generated by diffusion models	12-DoF Quadrotor Task; 7-DoF Robot Manipulator
EDMP[115]	2024	Motion Planning	Learning latent potentials and composing them as a cost function, enabling generalization to various motion constraints	Pybullet Simulator; MπNets
Luo et al. [116]	2024	Motion Planning	Learning latent potentials and composing them as a cost function, enabling generalization to various motion constraints	Maze2D;KUKA; Dual KUKA;Real-world
StructDiffusion [117]	2023	Task Planning; Rearrangement	Generating plausible object arrangement scenes using diffusion models to guide rearrangement tasks.	PyBullet Simulator; Real-world
DALL-E-Bot [118]	2023	Task Planning; Rearrangement	Applying web-scale network diffusion models to task planning for guiding robotic rearrangement tasks	Real-world
ReorientDiff [119]	2024	Task Planning; Rearrangement	Employing a diffusion model to sample intermediate poses from a joint representation, making rearrangement more precise	PyBullet Simulator; OMPL
SetItUp[120]	2024	Task Planning; Rearrangement	Harnessing LLMs to generate abstract spatial relationships between objects as constraint conditions, handling ambiguous instructions	Rearrangement on study desks, dining tables, and coffee tables
RPDiff[121]	2023	Task Planning; Rearrangement	Formulating rearrangement task as point cloud denoising process by learning geometric relationships between objects and scenes	PyBullet Simulator; Real-world
Diffusion-CCSP [122]	2023	Task Planning	Representing constraints as factor graphs by training diffusion models for every constraint and compositing them	Four Multi-Constraint Domains
GSC[123]	2023	Task Planning	Formulating robotic task as a skill chain, where diffusion models are trained for individual skills and assembled during inference.	Toy Domain; PyBullet Simulator
Planning as In-Painting [124]	2023	Task Planning	Formulating task planning as an image inpainting problem to jointly estimate target states and trajectories.	CompILE;Kuka Robot; ALFRED
Diffusion-EDFs [125]	2024	Task Planning	Developed a bi-equivariant diffusion model on the SE(3) manifold for task planning, which enhances sample efficiency.	SAPIEN; Real-world
SuSIE[126]	2024	Task Planning; Text-Image	A text-image diffusion model is employed to generate future frame from current frames and task descriptions for task guidance.	CALVIN; BridgeData V2
GENIMA[128]	2024	Task Planning; Text-Image	Joint target positions and rotation angles are rendered as colored spheres on the image to guide generation of joint position sequences.	CoppeliaSim; RLBench;Real-world
CoTDiffusion [129]	2024	Task Planning; Text-Image; Hierarchical	The diffusion model serves as a high-level visual planner to generate semantically aligned keyframes for visual guidance.	VIMA-BENCH
UniPi[130]	2023	Task Planning; Text-Video	Training a text-video diffusion model to generate highly realistic video plans to guide robotic action execution.	PDSketch;Cliport; Real-world
AVDC[131]	2024	Task Planning; Text-Video	Estimating optical flow between adjacent frames in synthetic videos, utilizing optical flow and initial frame depth to estimate actions.	Meta-World;iTHOR; Bridge;Real-world
HiP[132]	2023	Task Planning; Text-Video	Using LLMs to decompose language instructions into subgoal sequences, guiding the generation of video diffusion models.	Paint-Bl0ock;object-arrange; kitchen-tasks
VLP[133]	2023	Task Planning; Text-Video	Using VLMs to generate next-step action descriptions, combined with parallel hill climbing to enhance video reliability.	Language Table; Real-world
RoboDreamer [134]	2024	Task Planning; Text-Video	Parsing raw instructions into low-level language units and generating video plans based on these units.	RLBench;RT-1
This&That [135]	2024	Task Planning; Text-Video	Introducing additional gesture cues on initial frames to deliver targeted guidance for diffusion model	Bridge;IsaacGym
SkillDiffuser [136]	2024	Task Planning; Text-Video	Encoding skills into finite skill sets via Vector Quantization and utilizing skill abstractions as conditioning inputs for diffusion models.	LOReL Sawyer; Meta-World
Botteghi et al. [142]	2023	Motion Planning; Safety	Trainning a DDPM with CBF constraints for trajectory safety classification and security evaluation	Openai Gym
SafeDiffuser [143]	2025	Motion Planning; Safety	Integrating CBFs into conditional generation process of diffusion model to enforce generated trajectory safety	D4RL
Lee et al.[144]	2023	Motion Planning; Safety	Propose a new recovery gap metric to evaluate the feasibility of plans generated by diffusion models.	Maze2D;Gym MuJoCo; Block Stacking
LTLDoG[145]	2024	Motion Planning; Safety	Incorporating LTLf into the diffusion model to ensure generated trajectories satisfy specified static and temporal constraints	Maze2D;Push-T; Real-world

Table 4: Diffusion models in Task Planning and Reasoning

This method specifically calculates optical flow between consecutive frames in synthesized videos, establishing pixel-level dense correspondences. The system then combines this optical flow data with depth information from the initial frame. Through this integration, it determines the required robotic actions.

Subsequent research focuses on improving conditional inputs to diffusion models or incorporating additional information to enhance generation quality. For instance, HiP[131] employs two key components in robotic task planning. First, a video diffusion model performs visual planning of action sequences. Second, LLMs decompose language instructions into structured subgoal sequences that guide the video generation process. However, the subgoal sequences generated by LLMs remain static and fail to account for real-time robot states during task execution. In contrast, VLMs demonstrate dynamic reasoning capabilities by processing current state images as input. This enables continuous adaptation to environmental changes during operation. Therefore, VLP[132] employs VLMs to generate subsequent text-based actions. Using VLMs as heuristic functions, the parallel hill climbing algorithm conducts forward search in potential video sequence spaces, enhancing video generation reliability. Unlike text instruction generation methods, RoboDreamer[133] parses raw commands into low-level language units and generates video plans based on these units. This approach also accommodates multimodal inputs (e.g., target images and sketches), providing richer information for objective generation. This&That[134] improves video generation specificity by introducing additional gesture information on initial frames. This simplified task framework enables more reliable video outputs while reducing operational complexity. Distinct from all aforementioned methods, SkillDiffuser[135] learns discrete human-interpretable skill representations from visual observations and language instructions. It quantizes skills into a finite set through Vector Quantization, using skill abstraction as guidance for video diffusion models. Since the model does not directly utilize visual information and instructional data as guidance, it demonstrates enhanced generalization capability when handling diverse robot morphologies and ambiguous instructions.

3.5. Other Applications

In addition to the aforementioned applications, several studies have explored alternative pathways by applying diffusion models to novel research

directions in robotics. The widespread adoption of diffusion models has significantly promoted technological advancements across multiple dimensions.

The placement and stacking of objects involve many interactions. However, robotic systems typically lack prior knowledge about geometric properties of objects. To address this challenge, recent studies have employed diffusion models to infer reasonable object poses for robotic manipulation tasks. These models provide precise guidance for object placement operations. For instance, the 6-DoFusion method[136] utilizes diffusion models to predict interactive object poses. This approach specifically models object interactions by incorporating Signed Distance Field (SDF) values as supplementary inputs during training. The diffusion model processes both point cloud data and corresponding SDF values simultaneously. This dual-input strategy enables generation of stable 6-DoF poses that account for physical interaction constraints.

Many robotic manipulation tasks involve handling complex contact dynamics and geometric relationships. However, practical acquisition of tactile images under diverse object poses, positions, and contact conditions requires extensive experiment with precise control of environment. To address this challenge, Higuera et al.[137] developed a simulation-to-reality framework using diffusion models. Their approach generates realistic tactile images from simulated contact depth maps through an iterative denoising process. Experimental validation on braille reading tasks demonstrates significant advantages of the diffusion-based method. Compared with conventional techniques, the generated tactile images exhibit enhanced structural fidelity and finer textural details.

In addition to improving the ability to infer actions and trajectories, another effective approach to enhance task success rates involves addressing challenges through mechanical design. For specific tasks, customized grippers can be designed to reduce precision requirements in perception and control systems. The DGDM method[138] proposes a framework that decomposes manipulation tasks into sequential motion objectives called target interaction profiles. By comparing current and target interaction profiles, design objectives can be established to guide the geometric refinement of robotic fingers. However, this method employs oversimplified parameterization of actuators and stiffness properties, creating discrepancies between simulation results and real-world robotic performance. Future research could investigate more flexible parameterization approaches to bridge this implementation gap. DiffuseBot[139] focuses on soft robot design challenges. This

method first employs diffusion models to generate diverse 3D shapes as foundational robot geometries. The generated surface point clouds are then converted into robot-compatible representations incorporating material stiffness and actuator placement parameters. Finally, physical simulations are conducted to evaluate robot prototypes, enabling optimization of the diffusion model's input embeddings. Through this co-optimization process, the method demonstrates enhanced robotic performance and successfully generates high-efficiency soft robot designs.

4. Challenges and Limitations

After reviewing the relevant literature mentioned above, this section will focus on discussing the limitations and challenges. Some of these challenges are inherent to the field of robotics, and diffusion models show potential in addressing these issues. Based on these challenges and limitations, we provide prospects for future development in this field.

4.1. Safety Issues

For robots to achieve autonomy, they must first overcome various safety challenges. However, due to imperfect dynamic models, inadequate handling of sensor measurement noise, and insufficient characterization of operating environments, robotic systems often exhibit uncertain dynamic behaviors. Although robots possess limited knowledge of the real world, they must still make reasonable decisions while ensuring behavioral safety, particularly in human-robot interaction tasks. Consequently, safety remains an unavoidable challenge in robotics.

Traditional approaches to addressing safety and adaptability challenges have evolved through two primary methodologies. Control theory-based methods employ predefined dynamical models to ensure system safety, yet their reliance on accurate physical parameters restricts environmental adaptability. Conversely, RL techniques adopt data-driven exploration to achieve environmental flexibility, but inherently compromise safety guarantees during the learning phase.

Recent research efforts have developed safety-focused solutions using robust model predictive control. These approaches explicitly model uncertain system dynamics while maintaining state constraints. In parallel, RL methods now incorporate exploration boundaries or safety verification layers to generate more conservative policy updates. The pioneering work CPO[140]

established constrained policy optimization in the Constrained Markov Decision Process framework, ensuring policy improvements while satisfying safety limitations. Recent research has integrated constraint formulations with diffusion models to improve safety in robotic control tasks. Botteghi et al.[141] developed a framework combining Control Barrier Functions (CBFs) and DDPMs. The system employs three separate DDPM modules: trajectory generation aligned with system dynamics, expected return estimation, and safety classification using CBF constraints. A novel sampling strategy incorporates both value functions and safety classifiers as guidance signals, enabling simultaneous optimization of task performance and constraint satisfaction. This architecture modularizes trajectory planning, value prediction, and safety verification components typically coupled in RL, significantly enhancing operational flexibility while maintaining safety guarantees. SafeDiffuser[142] targets safety-critical trajectory planning in robotic navigation and manipulation tasks. This framework implements CBFs during the diffusion timesteps, embedding finite-time invariance properties into the denoising process to ensure probabilistic safety guarantees. Three specialized variants were developed to address different requirements: a constrained sampling module prevents local optima entrapment, an adaptive constraint relaxation mechanism balances safety-energy tradeoffs, and a hybrid optimization architecture preserves the diffusion model's generative performance. These innovations enable simultaneous safety enforcement and trajectory quality maintenance through time-dependent barrier function integration.

Lee et al.[143] proposes a novel metric called recovery margin to quantify the feasibility of trajectories generated by diffusion models. This metric evaluates trajectory quality by measuring the degree to which noise-corrupted plans can be restored to their original states through denoising operations. The approach effectively filters out physically unrealizable trajectories, thereby enhancing operational safety in robotic systems. Concurrently, LTLDoG[144] embeds finite linear temporal logic (LTLf) formulas into the diffusion process to ensure trajectory compliance with spatiotemporal constraints. The framework implements two strategies: using LTLf formulas as conditional guidance during diffusion sampling, and training a verification network to predict constraint satisfaction probabilities for guiding denoising trajectory generation. Experiments demonstrate that LTLDoG successfully generates feasible trajectories meeting temporal logic requirements in both simulation environments and quadruped robot navigation tasks.

Future research should focus on balancing safety requirements in robotic

robotics with preserving generative capability of diffusion model. This equilibrium presents a critical challenge for practical deployments. High-dimensional long-horizon tasks particularly demand solutions for managing multiple concurrent constraints. Developing more efficient mechanisms to meet the constraints could significantly improve mission success rates.

4.2. Real-Time Inference and Model Size

While diffusion models show broad applications in robotic trajectory planning, motion generation, and policy synthesis, their practical implementation faces two critical limitations: substantial parameter sizes and heavy data dependency. These constraints particularly hinder real-time inference in complex tasks.To address these challenges, two strategic directions emerge. First, model simplification through architectural optimization could be pursued. Implementing model pruning techniques may effectively eliminate non-critical parameters. Concurrently, knowledge distillation methods could transfer knowledge from large diffusion models to compact versions without significant performance loss. Second, accelerated sampling methods specifically designed for diffusion processes should be integrated. Recent advancements like DPM-Solver++[145] and LCM-LoRA[146] demonstrate the feasibility of generating high-quality outputs with minimal diffusion steps. Combining these acceleration frameworks with policy generation pipelines could achieve real-time capability while preserving generative performance.

Beyond algorithmic improvements, hardware optimization offers complementary acceleration opportunities. Implementing advanced GPU or TPU architectures can significantly boost computational efficiency for diffusion model operations. Meanwhile, edge computing implementations provide practical solutions for real-time applications. By deploying computational workloads on edge devices closer to data sources, this approach minimizes communication latency with cloud servers.

4.3. Simulation to the Real World Gap

Acquiring real-world robotic datasets remains challenging due to high costs and technical constraints. Most current studies conduct experiments in simulated environments. This reliance creates the Sim2Real gap that significantly hinders embodied AI development.

The primary contributor to Sim2Real gap stems from disparities in physical modeling and environmental complexity. Current simulation platforms implement simplifications in dynamic properties, material characteristics,

and scene configurations. These approximations render simulation-optimized approaches ineffective in physical deployments. Furthermore, real-world sensors exhibit distinct noise profiles that disrupt robotic perception systems. Additionally, computational latency in physical systems significantly exceeds simulation predictions, demanding more robust control algorithms to maintain operational stability.

The research community has witnessed a proliferation of novel conceptual frameworks emerging in recent years to systematically address these challenges. A predominant strategy employs domain randomization[44][45] in simulation environments. This approach systematically varies physical parameters including friction coefficients and material densities, while concurrently modifying environmental factors such as illumination spectra and sensor noise profiles. Though domain randomization improves cross-domain performance compared to deterministic simulation paradigms, substantial gaps in task success rates persist when transitioning to unstructured environments.

Another approach is digital twin technology. A digital twin is a digital replica of a physical system. It can mirror the states of physical system in real time through bidirectional data exchange, which facilitates system optimization and control. Current applications span space robots, medical robots, soft robots and industrial robots[147]. Digital twins help bridge the Sim2Real gap in specific environments but require high implementation costs. Recently, the ACDC framework [148] has attracted research interest. This framework automatically generates virtual environments which called digital cousins from single RGB images. These synthetic environments resemble real-world scenes while maintaining lower generation costs. Digital cousins preserve geometric and semantic similarities to real counterparts without precise modeling. This design reduces generation costs. The diversified synthetic environments also enhance algorithmic robustness. However, this method still faces limitations. Its effectiveness depends on the diversity of underlying asset datasets and in-domain data availability. Consequently, real-world deployment remains challenging.

Future research may focus on leveraging diffusion models to create more realistic simulations. Improving simulation fidelity through such generative models represents a promising direction in this field.

4.4. Datasets and Unified Benchmarks

The scarcity of datasets remains a significant barrier in robot learning research. Although diffusion models can scale up robot manipulation datasets to some extent, they cannot generate realistic robot datasets with arbitrary scenes, objects, and diverse manipulation strategies. Researchers have begun developing large-scale datasets for robotics, such as Open X-embodiment[149], Robomind[150], and AgiBot World[151]. These datasets cover multiple robot platforms, tasks, and environments. However, several challenges persist. First, data structures lack standardization, and input modalities only partially include image, 3D vision, text, or tactile signals. In addition, incompatible formats between multi-robot datasets complicate data processing and loading. As computing power continues to grow and data accumulates, creating general-purpose pretraining models for robots has become an important research direction. This trend highlights the need for unified large-scale multi-robot datasets with standardized structures and formats, which remains a critical development goal in robotics.

Moreover, the limited availability of public benchmarks in robotics significantly hampers fair performance evaluations. Current benchmarks primarily focus on single-arm robots, while critical platforms such as dual-arm manipulators, dexterous hands, and quadruped robots still lack dedicated evaluation frameworks. Standardized assessment metrics remain underdeveloped in this field. Although task success rates serve as a common evaluation criterion, they fail to quantify crucial capabilities including cross-environment adaptability, viewpoint generalization, task transferability, and safety. To address these gaps, expanding simulation-based benchmarks and accelerating real-world benchmark development represent essential directions for robotic research. Establishing unified evaluation protocols across diverse robotic platforms and tasks will better support technological advancements in this domain.

4.5. Embodied Foundation Models

Foundation models trained on massive and diverse datasets demonstrate strong generalization capabilities and cross-task adaptability. While foundation models have achieved relative maturity in NLP and CV domains, significant developmental potential remains in robotics applications.

Recent research efforts have emerged to develop cross-environment, cross-task, cross-embodiment embodied foundation models. The primary challenge in developing such models arises from robotic data heterogeneity. Divergent

platform specifications and control signals across experimental setups result in significant action label discrepancies, impeding data sharing across different sources. RT-X series[149], Octo[152], and Open-VLA[153] have adopted discretization and normalization techniques to achieve action-space equivalence across heterogeneous platforms. They employ heterogeneous 7-DoF robotic datasets to enhance policy generalization. However, this approach fails to ensure action-space consistency and interpretability, with persistent risks of conflicting maneuvers. CrossFormer[154], RDT[108], $\pi 0$[155], and Yang et al.[156] aggregate heterogeneous action spaces to achieve cross-robot generalization while enhancing action-space interpretability, albeit requiring significant manual effort for action-space construction. UniAct[157] utilizes VLMs as action extractors to learn fundamental action primitives from diverse single-arm robotic datasets. These learned representations are then adapted to various robotic platforms and control modalities through multiple dedicated decoder heads.

While existing approaches partially mitigate data heterogeneity limitations, fundamental challenges persist in ensuring action-space interpretability. Future research directions should prioritize developing interpretable cross-embodiment action spaces and creating foundational models capable of cross-embodiment control through heterogeneous control signals.

5. Conclusion

This paper systematically reviews the application of diffusion models in robotic systems and comprehensively analyzes current research advancements. We first outlines the fundamental principles of diffusion models and their recent technological progress. Subsequently, the paper categorizes the functional roles of diffusion models across five primary domains: scaling up robotics data, RL, IL, task planning and reasoning and other applications. These applications effectively address persistent challenges in robotics such as insufficient training data, limited action modalities, and weak generalization capabilities. The concluding section identifies both existing challenges in robotics and limitations in applying diffusion models, particularly regarding safety concerns, real-time processing requirements, simulation-to-reality gaps, Embodied foundation models, and the lack of standardized benchmarks. These unresolved issues present crucial research opportunities for future investigations. The study emphasizes that advanced generative models

like diffusion architectures possess significant potential to accelerate robotic innovation.

References

[1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 1 (2) (2022) 3.

[3] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, A. Ramesh, Video generation models as world simulators (2024).

[4] H. Wang, D. Meger, Robotic object manipulation with full-trajectory gan-based imitation learning, in: 2021 18th Conference on Robots and Vision (CRV), IEEE, 2021, pp. 57–63.

[5] H. Zhan, F. Tao, Y. Cao, Human-guided robot behavior learning: A gan-assisted preference-based reinforcement learning approach, IEEE Robotics and Automation Letters 6 (2) (2021) 3545–3552.

[6] V. Huang, T. Ley, M. Vlachou-Konchylaki, W. Hu, Enhanced experience replay generation for efficient reinforcement learning, arXiv preprint arXiv:1705.08245 (2017).

[7] B. Imre, An investigation of generative replay in deep reinforcement learning, B.S. thesis, University of Twente (2021).

[8] H.-S. Moon, J. Seo, Observation of human response to a robotic guide using a variational autoencoder, in: 2019 Third IEEE International Conference on Robotic Computing (IRC), IEEE, 2019, pp. 258–261.

[9] T. Ubukata, J. Li, K. Tei, Diffusion model for planning: A systematic literature review, arXiv preprint arXiv:2408.10266 (2024).

[10] K. Zhang, P. Yun, J. Cen, J. Cai, D. Zhu, H. Yuan, C. Zhao, T. Feng, M. Y. Wang, Q. Chen, et al., Generative artificial intelligence in robotic manipulation: A survey, arXiv preprint arXiv:2503.03464 (2025).

[11] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in neural information processing systems 33 (2020) 6840–6851.

[12] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International conference on machine learning, PMLR, 2015, pp. 2256–2265.

[13] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, Advances in neural information processing systems 32 (2019).

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, in: International Conference on Learning Representations, 2021.

[15] L. Liu, Y. Ren, Z. Lin, Z. Zhao, Pseudo numerical methods for diffusion models on manifolds, arXiv preprint arXiv:2202.09778 (2022).

[16] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu, Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, Advances in Neural Information Processing Systems 35 (2022) 5775–5787.

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

[18] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Advances in neural information processing systems 34 (2021) 8780–8794.

[19] J. Ho, T. Salimans, Classifier-free diffusion guidance, arXiv preprint arXiv:2207.12598 (2022).

[20] T. Karras, M. Aittala, T. Aila, S. Laine, Elucidating the design space of diffusion-based generative models, Advances in neural information processing systems 35 (2022) 26565–26577.

[21] A. Vahdat, K. Kreis, J. Kautz, Score-based generative modeling in latent space, Advances in neural information processing systems 34 (2021) 11287–11302.

[22] D. Kim, B. Na, S. J. Kwon, D. Lee, W. Kang, I.-c. Moon, Maximum likelihood training of parametrized diffusion model (2021).

[23] Y. Xu, Z. Liu, M. Tegmark, T. Jaakkola, Poisson flow generative models, Advances in Neural Information Processing Systems 35 (2022) 16782–16795.

[24] A. Bansal, E. Borgnia, H.-M. Chu, J. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, T. Goldstein, Cold diffusion: Inverting arbitrary image transforms without noise, Advances in Neural Information Processing Systems 36 (2024).

[25] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, M. Le, Flow matching for generative modeling, in: The Eleventh International Conference on Learning Representations, 2023.

[26] R. G. Lopes, S. Fenu, T. Starner, Data-free knowledge distillation for deep neural networks, arXiv preprint arXiv:1710.07535 (2017).

[27] T. Salimans, J. Ho, Progressive distillation for fast sampling of diffusion models, in: International Conference on Learning Representations, 2022.

[28] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, T. Salimans, On distillation of guided diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14297–14306.

[29] E. Luhman, T. Luhman, Knowledge distillation in iterative generative models for improved sampling speed, arXiv preprint arXiv:2101.02388 (2021).

[30] Y. Song, P. Dhariwal, M. Chen, I. Sutskever, Consistency models, in: International Conference on Machine Learning, PMLR, 2023, pp. 32211–32252.

[31] D. Berthelot, A. Autef, J. Lin, D. A. Yap, S. Zhai, S. Hu, D. Zheng, W. Talbott, E. Gu, Tract: Denoising diffusion models with transitive closure time-distillation, arXiv preprint arXiv:2303.04248 (2023).

[32] X. Liu, C. Gong, Q. Liu, Flow straight and fast: Learning to generate and transfer data with rectified flow, in: The Eleventh International Conference on Learning Representations (ICLR), 2023.

[33] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations, 2021.
URL https://openreview.net/forum?id=St1giarCHLP

[34] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman, I. Mitliagkas, Gotta go fast when generating data with score-based models, arXiv preprint arXiv:2105.14080 (2021).

[35] Y. Xu, M. Deng, X. Cheng, Y. Tian, Z. Liu, T. Jaakkola, Restart sampling for improving generative processes, Advances in Neural Information Processing Systems 36 (2023) 76806–76838.

[36] Y. Song, C. Durkan, I. Murray, S. Ermon, Maximum likelihood training of score-based diffusion models, Advances in neural information processing systems 34 (2021) 1415–1428.

[37] D. Kingma, T. Salimans, B. Poole, J. Ho, Variational diffusion models, Advances in neural information processing systems 34 (2021) 21696–21707.

[38] C.-W. Huang, J. H. Lim, A. C. Courville, A variational perspective on diffusion-based generative models and score matching, Advances in Neural Information Processing Systems 34 (2021) 22863–22876.

[39] A. Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International conference on machine learning, PMLR, 2021, pp. 8162–8171.

[40] C. Lu, K. Zheng, F. Bao, J. Chen, C. Li, J. Zhu, Maximum likelihood training for score-based diffusion odes by high order denoising score matching, in: International Conference on Machine Learning, PMLR, 2022, pp. 14429–14460.

[41] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Advances in neural information processing systems 35 (2022) 36479–36494.

[42] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, L. Fan, Vima: General robot manipulation with multimodal prompts, arXiv preprint arXiv:2210.03094 2 (3) (2022) 6.

[43] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al., Rt-1: Robotics transformer for real-world control at scale, arXiv preprint arXiv:2212.06817 (2022).

[44] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, 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), IEEE, 2017, pp. 23–30.

[45] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, L. Paull, Active domain randomization, in: Conference on Robot Learning, PMLR, 2020, pp. 1162–1176.

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

[47] Z. Chen, S. Kiami, A. Gupta, V. Kumar, Genaug: Retargeting behaviors to unseen situations via generative augmentation, arXiv preprint arXiv:2302.06671 (2023).

[48] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, et al., Scaling robot learning with semantically imagined experience, arXiv preprint arXiv:2302.11550 (2023).

[49] X. Zhang, M. Chang, P. Kumar, S. Gupta, Diffusion meets dagger: Supercharging eye-in-hand imitation learning, in: Robotics science and systems, Robotics science and systems, 2024.

[50] L. Y. Chen, C. Xu, K. Dharmarajan, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, K. Goldberg, Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning, in: 8th Annual Conference on Robot Learning, 2024.

[51] A. Kirillov, E. Mintun, N. Ravi, et al., Segment anything proceedings of the ieee, in: CVF International Conference on Computer Vision (ICCV), IEEE, 2023, pp. 4015–4026.

[52] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, M.-M. Cheng, Towards an end-to-end framework for flow-guided video inpainting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17562–17571.

[53] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun, et al., Zeronvs: Zero-shot 360-degree view synthesis from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9420–9429.

[54] C. Lu, P. Ball, Y. W. Teh, J. Parker-Holder, Synthetic experience replay, Advances in Neural Information Processing Systems 36 (2024).

[55] P. Katara, Z. Xian, K. Fragkiadaki, Gen2sim: Scaling up robot learning in simulation with generative models, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 6672–6679.

[56] R. OpenAI, Gpt-4 technical report. arxiv 2303.08774, View in Article 2 (5) (2023).

[57] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, C. Vondrick, Zero-1-to-3: Zero-shot one image to 3d object, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 9298–9309.

[58] Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, C. Gan, Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, in: Forty-first International Conference on Machine Learning, 2024.

[59] M. Janner, Y. Du, J. Tenenbaum, S. Levine, Planning with diffusion for flexible behavior synthesis, in: Proceedings of the 39th International Conference on Machine Learning, 2022, pp. 9902–9915.

[60] Z. Wang, J. J. Hunt, M. Zhou, Diffusion policies as an expressive policy class for offline reinforcement learning, in: The Eleventh International Conference on Learning Representations, 2023.

[61] H. Chen, C. Lu, C. Ying, H. Su, J. Zhu, Offline reinforcement learning via high-fidelity generative behavior modeling, in: The Eleventh International Conference on Learning Representations, 2023.

[62] L. He, L. Shen, L. Zhang, J. Tan, X. Wang, Diffcps: Diffusion model based constrained policy search for offline reinforcement learning, arXiv preprint arXiv:2310.05333 (2023).

[63] S. E. Ada, E. Oztop, E. Ugur, Diffusion policies for out-of-distribution generalization in offline reinforcement learning, IEEE Robotics and Automation Letters (2024).

[64] J. Brehmer, J. Bose, P. De Haan, T. S. Cohen, Edgi: Equivariant diffusion for planning with embodied agents, Advances in Neural Information Processing Systems 36 (2024).

[65] P. Hansen-Estruch, I. Kostrikov, M. Janner, J. G. Kuba, S. Levine, Idql: Implicit q-learning as an actor-critic method with diffusion policies, arXiv preprint arXiv:2304.10573 (2023).

[66] A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, P. Agrawal, Is conditional generative modeling all you need for decision making?, in: The Eleventh International Conference on Learning Representations, 2023.

[67] C. Lu, H. Chen, J. Chen, H. Su, C. Li, J. Zhu, Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning, in: International Conference on Machine Learning, PMLR, 2023, pp. 22825–22855.

[68] S. Venkatraman, S. Khaitan, R. T. Akella, J. Dolan, J. Schneider, G. Berseth, Reasoning with latent diffusion in offline reinforcement

learning, in: The Twelfth International Conference on Learning Representations, 2024.

[69] S. Fujimoto, D. Meger, D. Precup, Off-policy deep reinforcement learning without exploration, in: International conference on machine learning, PMLR, 2019, pp. 2052–2062.

[70] E. Mitchell, R. Rafailov, X. B. Peng, S. Levine, C. Finn, Offline meta-reinforcement learning with advantage weighting, in: International Conference on Machine Learning, PMLR, 2021, pp. 7780–7791.

[71] J. Li, Q. Vuong, S. Liu, M. Liu, K. Ciosek, H. Christensen, H. Su, Multi-task batch reinforcement learning with metric learning, Advances in neural information processing systems 33 (2020) 6197–6210.

[72] H. Kurunathan, H. Huang, K. Li, W. Ni, E. Hossain, Machine learning-aided operations and communications of unmanned aerial vehicles: A contemporary survey, IEEE Communications Surveys & Tutorials (2023).

[73] F. Ni, J. Hao, Y. Mu, Y. Yuan, Y. Zheng, B. Wang, Z. Liang, Metadiffuser: Diffusion model as conditional planner for offline meta-rl, in: International Conference on Machine Learning, PMLR, 2023, pp. 26087–26105.

[74] Z. Liang, Y. Mu, M. Ding, F. Ni, M. Tomizuka, P. Luo, Adaptdiffuser: Diffusion models as adaptive self-evolving planners, in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023, pp. 20725–20745.

[75] Z. Dong, J. Hao, Y. Yuan, F. Ni, Y. Wang, P. Li, Y. Zheng, Diffuserlite: Towards real-time diffusion planning, arXiv preprint arXiv:2401.15443 (2024).

[76] C. Chen, F. Deng, K. Kawaguchi, C. Gulcehre, S. Ahn, Simple hierarchical planning with diffusion, in: The Twelfth International Conference on Learning Representations, 2024.

[77] W. Li, X. Wang, B. Jin, H. Zha, Hierarchical diffusion for offline decision making, in: International Conference on Machine Learning, PMLR, 2023, pp. 20035–20064.

[78] Z. Zhu, M. Liu, L. Mao, B. Kang, M. Xu, Y. Yu, S. Ermon, W. Zhang, Madiff: Offline multi-agent learning with diffusion models, arXiv preprint arXiv:2305.17330 (2023).

[79] Z. Li, L. Pan, L. Huang, Beyond conservatism: Diffusion policies in offline multi-agent reinforcement learning, arXiv preprint arXiv:2307.01472 (2023).

[80] L. Yang, Z. Huang, F. Lei, Y. Zhong, Y. Yang, C. Fang, S. Wen, B. Zhou, Z. Lin, Policy representation via diffusion probability model for reinforcement learning, arXiv preprint arXiv:2305.13122 (2023).

[81] Y. Chen, H. Li, D. Zhao, Boosting continuous control with consistency policy, in: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, 2024, pp. 335–344.

[82] M. Rigter, J. Yamada, I. Posner, World models via policy-guided trajectory diffusion, Transactions on Machine Learning Research (2024).

[83] Y. Wang, L. Wang, Y. Jiang, W. Zou, T. Liu, X. Song, W. Wang, L. Xiao, J. Wu, J. Duan, et al., Diffusion actor-critic with entropy regulator, arXiv preprint arXiv:2405.15177 (2024).

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

[85] S. Hegde, S. Batra, K. Zentner, G. Sukhatme, Generating behaviorally diverse policies with latent diffusion models, in: Advances in Neural Information Processing Systems, Vol. 36, Curran Associates, Inc., 2023, pp. 7541–7554.

[86] L. Chen, S. Bahl, D. Pathak, Playfusion: Skill acquisition via diffusion from language-annotated play, in: Conference on Robot Learning, PMLR, 2023, pp. 2012–2029.

[87] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann, S. Devlin, Imitating human behaviour with diffusion models, in: The Eleventh International Conference on Learning Representations, 2023.

[88] M. Reuss, M. Li, X. Jia, R. Lioutikov, Goal-conditioned imitation learning using score-based diffusion policies, in: Robotics: Science and Systems, 2023.

[89] Z. Xian, N. Gkanatsios, T. Gervet, T.-W. Ke, K. Fragkiadaki, Chained-diffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation, in: 7th Annual Conference on Robot Learning, 2023.

[90] E. Ng, Z. Liu, M. Kennedy, Diffusion co-policy for synergistic human-robot collaborative tasks, IEEE Robotics and Automation Letters (2023).

[91] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, S. Song, Diffusion policy: Visuomotor policy learning via action diffusion, The International Journal of Robotics Research (2023) 02783649241273668.

[92] A. Prasad, K. Lin, J. Wu, L. Zhou, J. Bohg, Consistency policy: Accelerated visuomotor policies via consistency distillation, arXiv preprint arXiv:2405.07503 (2024).

[93] H. Zhou, D. Blessing, G. Li, O. Celik, X. Jia, G. Neumann, R. Lioutikov, Variational distillation of diffusion policies into mixture of experts, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[94] S. H. Høeg, Y. Du, O. Egeland, Streaming diffusion policy: Fast policy synthesis with variable noise diffusion models (2024). arXiv:2406.04806.

[95] K. Chen, E. Lim, K. Lin, Y. Chen, H. Soh, Don't start from scratch: Behavioral refinement via interpolant-based policy diffusion, arXiv preprint arXiv:2402.16075 (2024).

[96] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, R. Platt, Equivariant diffusion policy, in: 8th Annual Conference on Robot Learning, 2024.

[97] X. Li, V. Belagali, J. Shang, M. S. Ryoo, Crossway diffusion: Improving diffusion-based visuomotor policy via self-supervised learning, in: 2024

IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 16841–16849.

[98] X. Ma, S. Patidar, I. Haughton, S. James, Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18081–18090.

[99] C. Hao, K. Lin, S. Luo, H. Soh, Language-guided manipulation with diffusion policies and constrained inpainting, arXiv preprint arXiv:2406.09767 (2024).

[100] H. Bharadhwaj, R. Mottaghi, A. Gupta, S. Tulsiani, Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation, in: European Conference on Computer Vision, Springer, 2024, pp. 306–324.

[101] M. Xu, Z. Xu, C. Chi, M. Veloso, S. Song, Xskill: Cross embodiment skill discovery, in: Conference on Robot Learning, PMLR, 2023, pp. 3536–3555.

[102] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, H. Xu, 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations, in: ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation, 2024.

[103] T.-W. Ke, N. Gkanatsios, K. Fragkiadaki, 3d diffuser actor: Policy diffusion with 3d scene representations, in: ICRA 2024 Workshop—Back to the Future: Robot Learning Going Probabilistic, 2024.

[104] G. Yan, Y.-H. Wu, X. Wang, Dnact: Diffusion guided multi-task 3d policy learning, arXiv preprint arXiv:2403.04115 (2024).

[105] V. Vosylius, Y. Seo, J. Uruç, S. James, Render and diffuse: Aligning image and action spaces for diffusion-based behaviour cloning, arXiv preprint arXiv:2405.18196 (2024).

[106] A. Sridhar, D. Shah, C. Glossop, S. Levine, Nomad: Goal masked diffusion policies for navigation and exploration, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 63–70.

[107] T. Yoneda, L. Sun, G. Yang, B. C. Stadie, M. R. Walter, To the noise and back: Diffusion for shared autonomy, in: Robotics: Science and Systems, 2023.

[108] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, J. Zhu, Rdt-1b: a diffusion foundation model for bimanual manipulation, arXiv preprint arXiv:2410.07864 (2024).

[109] J. Carvalho, M. Baierl, J. Urain, J. Peters, Conditioned score-based models for learning collision-free trajectory generation, in: NeurIPS 2022 Workshop on Score-Based Methods, 2022.

[110] J. Urain, N. Funk, J. Peters, G. Chalvatzaki, Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 5923–5930.

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

[112] X. Fang, C. R. Garrett, C. Eppner, T. Lozano-Pérez, L. P. Kaelbling, D. Fox, Dimsam: Diffusion models as samplers for task and motion planning under partial observability, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2024, pp. 1412–1419.

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

[114] K. Saha, V. Mandadi, J. Reddy, A. Srikanth, A. Agarwal, B. Sen, A. Singh, M. Krishna, Edmp: Ensemble-of-costs-guided diffusion for motion planning, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 10351–10358.

[115] Y. Luo, C. Sun, J. B. Tenenbaum, Y. Du, Potential based diffusion motion planning, in: International Conference on Machine Learning, PMLR, 2024, pp. 33486–33510.

[116] W. Liu, Y. Du, T. Hermans, S. Chernova, C. Paxton, Structdiffusion: Language-guided creation of physically-valid structures using unseen objects, in: Robotics: Science and Systems, 2023.

[117] I. Kapelyukh, V. Vosylius, E. Johns, Dall-e-bot: Introducing web-scale diffusion models to robotics, IEEE Robotics and Automation Letters 8 (7) (2023) 3956–3963.

[118] U. A. Mishra, Y. Chen, Reorientdiff: Diffusion model based reorientation for object manipulation, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 10867–10873.

[119] Y. Xu, J. Mao, Y. Du, T. Lozano-Pérez, L. P. Kaelbling, D. Hsu, " set it up!": Functional object arrangement with compositional generative models, CoRR (2024).

[120] A. Simeonov, A. Goyal, L. Manuelli, Y.-C. Lin, A. Sarmiento, A. R. Garcia, P. Agrawal, D. Fox, Shelving, stacking, hanging: Relational pose diffusion for multi-modal rearrangement, in: 7th Annual Conference on Robot Learning, 2023.

[121] Z. Yang, J. Mao, Y. Du, J. Wu, J. B. Tenenbaum, T. Lozano-Pérez, L. P. Kaelbling, Compositional diffusion-based continuous constraint solvers, in: Conference on Robot Learning, PMLR, 2023, pp. 3242–3265.

[122] U. A. Mishra, S. Xue, Y. Chen, D. Xu, Generative skill chaining: Long-horizon skill planning with diffusion models, in: Conference on Robot Learning, PMLR, 2023, pp. 2905–2925.

[123] C.-F. Yang, H. Xu, T.-L. Wu, X. Gao, K.-W. Chang, F. Gao, Planning as in-painting: A diffusion-based embodied task planning framework for environments under uncertainty, CoRR (2023).

[124] H. Ryu, J. Kim, H. An, J. Chang, J. Seo, T. Kim, Y. Kim, C. Hwang, J. Choi, R. Horowitz, 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, 2024, pp. 18007–18018.

[125] K. Black, M. Nakamoto, P. Atreya, H. R. Walke, C. Finn, A. Kumar, S. Levine, Zero-shot robotic manipulation with pre-trained image-editing diffusion models, in: The Twelfth International Conference on Learning Representations, 2024.

[126] T. Brooks, A. Holynski, A. A. Efros, Instructpix2pix: Learning to follow image editing instructions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 18392–18402.

[127] M. Shridhar, Y. L. Lo, S. James, Generative image as action models, in: 8th Annual Conference on Robot Learning, 2024.

[128] F. Ni, J. Hao, S. Wu, L. Kou, J. Liu, Y. Zheng, B. Wang, Y. Zhuang, Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13991–14000.

[129] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, P. Abbeel, Learning universal policies via text-guided video generation, Advances in neural information processing systems 36 (2023) 9156–9172.

[130] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, J. B. Tenenbaum, Learning to act from actionless videos through dense correspondences, in: The Twelfth International Conference on Learning Representations, 2024.

[131] A. Ajay, S. Han, Y. Du, S. Li, A. Gupta, T. Jaakkola, J. Tenenbaum, L. Kaelbling, A. Srivastava, P. Agrawal, Compositional foundation models for hierarchical planning, Advances in Neural Information Processing Systems 36 (2023) 22304–22325.

[132] Y. Du, M. Yang, P. Florence, F. Xia, A. Wahid, B. Ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum, et al., Video language planning, arXiv preprint arXiv:2310.10625 (2023).

[133] S. Zhou, Y. Du, J. Chen, Y. Li, D.-Y. Yeung, C. Gan, Robodreamer: learning compositional world models for robot imagination, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 61885–61896.

[134] B. Wang, N. Sridhar, C. Feng, M. Van der Merwe, A. Fishman, N. Fazeli, J. J. Park, This&that: Language-gesture controlled video generation for robot planning, arXiv preprint arXiv:2407.05530 (2024).

[135] Z. Liang, Y. Mu, H. Ma, M. Tomizuka, M. Ding, P. Luo, 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, 2024, pp. 16467–16476.

[136] T. Yoneda, T. Jiang, G. Shakhnarovich, M. R. Walter, 6-dof stability field via diffusion models, arXiv preprint arXiv:2310.17649 (2023).

[137] C. Higuera, B. Boots, M. Mukadam, Learning to read braille: Bridging the tactile reality gap with diffusion models, arXiv preprint arXiv:2304.01182 (2023).

[138] X. Xu, H. Ha, S. Song, Dynamics-guided diffusion model for robot manipulator design, arXiv preprint arXiv:2402.15038 (2024).

[139] T.-H. J. Wang, J. Zheng, P. Ma, Y. Du, B. Kim, A. Spielberg, J. Tenenbaum, C. Gan, D. Rus, Diffusebot: Breeding soft robots with physics-augmented generative diffusion models, Advances in Neural Information Processing Systems 36 (2023) 44398–44423.

[140] J. Achiam, D. Held, A. Tamar, P. Abbeel, Constrained policy optimization, in: International conference on machine learning, PMLR, 2017, pp. 22–31.

[141] N. Botteghi, F. Califano, M. Poel, C. Brune, Trajectory generation, control, and safety with denoising diffusion probabilistic models, in: ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems, 2023.

[142] W. Xiao, T.-H. Wang, C. Gan, R. Hasani, M. Lechner, D. Rus, Safediffuser: Safe planning with diffusion probabilistic models, in: The Thirteenth International Conference on Learning Representations, 2025.

[143] K. Lee, S. Kim, J. Choi, Refining diffusion planner for reliable behavior synthesis by automatic detection of infeasible plans, Advances in Neural Information Processing Systems 36 (2023) 24223–24246.

[144] Z. Feng, H. Luan, P. Goyal, H. Soh, Ltldog: Satisfying temporally-extended symbolic constraints for safe diffusion-based planning, IEEE Robotics and Automation Letters (2024).

[145] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu, Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, arXiv preprint arXiv:2211.01095 (2022).

[146] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, H. Zhao, Lcm-lora: A universal stable-diffusion acceleration module, arXiv preprint arXiv:2311.05556 (2023).

[147] A. Mazumder, M. F. Sahed, Z. Tasneem, P. Das, F. R. Badal, M. F. Ali, M. H. Ahamed, S. H. Abhi, S. K. Sarker, S. K. Das, et al., Towards next generation digital twin in robotics: Trends, scopes, challenges, and future, Heliyon 9 (2) (2023).

[148] T. Dai, J. Wong, Y. Jiang, C. Wang, C. Gokmen, R. Zhang, J. Wu, L. Fei-Fei, Automated creation of digital cousins for robust policy learning, arXiv preprint arXiv:2410.07408 (2024).

[149] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al., Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 6892–6903.

[150] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, et al., Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation, arXiv preprint arXiv:2412.13877 (2024).

[151] AgiBot-World-Contributors, Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Huang, S. Jiang, Y. Jiang, C. Jing, H. Li, J. Li, C. Liu, Y. Liu, Y. Lu, J. Luo, P. Luo, Y. Mu, Y. Niu, Y. Pan, J. Pang, Y. Qiao, G. Ren, C. Ruan, J. Shan, Y. Shen, C. Shi, M. Shi, M. Shi, C. Sima, J. Song, H. Wang, W. Wang, D. Wei, C. Xie, G. Xu, J. Yan, C. Yang, L. Yang, S. Yang, M. Yao, J. Zeng, C. Zhang, Q. Zhang, B. Zhao, C. Zhao, J. Zhao, J. Zhu, AgiBot World Colosseo:

A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems, arXiv preprint arXiv:2503.06669 (2025).

[152] O. Mees, D. Ghosh, K. Pertsch, K. Black, H. R. Walke, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, D. Sadigh, C. Finn, S. Levine, Octo: An open-source generalist robot policy, in: First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024.

[153] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, C. Finn, OpenVLA: An open-source vision-language-action model, in: 8th Annual Conference on Robot Learning, 2024.

[154] R. Doshi, H. R. Walke, O. Mees, S. Dasari, S. Levine, Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation, in: 8th Annual Conference on Robot Learning, 2024.

[155] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al., $\pi 0$: A vision-language-action flow model for general robot control, 2024, URL https://arxiv.org/abs/2410.24164 (2024).

[156] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, S. Levine, Pushing the limits of cross-embodiment learning for manipulation and navigation, arXiv preprint arXiv:2402.19432 (2024).

[157] J. Zheng, J. Li, D. Liu, Y. Zheng, Z. Wang, Z. Ou, Y. Liu, J. Liu, Y.-Q. Zhang, X. Zhan, Universal actions for enhanced embodied foundation models, arXiv preprint arXiv:2501.10105 (2025).