

Diffusion Models for Reinforcement Learning: A Survey

Zhengbang Zhu, Hanye Zhao, Haoran He, Yichao Zhong
Shenyu Zhang, Yong Yu, Weinan Zhang

Shanghai Jiao Tong University
{zhengbangzhu, wnzhang}@sjtu.edu.cn

Abstract

Diffusion models have emerged as a prominent class of generative models, surpassing previous methods regarding sample quality and training stability. Recent works have shown the advantages of diffusion models in improving reinforcement learning (RL) solutions, including as trajectory planners, expressive policy classes, data synthesizers, etc. This survey aims to provide an overview of the advancements in this emerging field and hopes to inspire new avenues of research. First, we examine several challenges encountered by current RL algorithms. Then, we present a taxonomy of existing methods based on the roles played by diffusion models in RL and explore how the existing challenges are addressed. We further outline successful applications of diffusion models in various RL-related tasks while discussing the limitations of current approaches. Finally, we conclude the survey and offer insights into future research directions, focusing on enhancing model performance and applying diffusion models to broader tasks. We are actively maintaining a GitHub repository for papers and other related resources in applying diffusion models in RL¹.

1 Introduction

Diffusion models have emerged as a powerful class of generative models, garnering significant attention in recent years. These models employ a denoising framework that can effectively reverse a multistep noising process to generate new data [Song *et al.*, 2021]. In contrast to earlier generative models such as Variational Autoencoders (VAE) [Kingma and Welling, 2013] and Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014], diffusion models exhibit superior capabilities in generating high-quality samples and demonstrate enhanced training stability. Consequently, they have made remarkable strides and achieved substantial success in diverse domains including computer vision [Ho *et al.*, 2020; Lugmayr *et al.*, 2022; Luo and Hu, 2021], natural language processing [Austin *et al.*, 2021;

Li *et al.*, 2022], audio generation [Lee and Han, 2021; Kong *et al.*, 2020], and drug discovery [Xu *et al.*, 2022; Schneuing *et al.*, 2022], *etc.*

Reinforcement learning (RL) [Sutton and Barto, 2018] focuses on training agents to solve sequential decision-making tasks by maximizing cumulative rewards. While RL has achieved remarkable successes in various domains [Kober *et al.*, 2013; Kiran *et al.*, 2021], there are some long-standing challenges. Specifically, despite the considerable attention garnered by offline RL for overcoming low sample efficiency issue in online RL [Kumar *et al.*, 2020; Fujimoto and Gu, 2021], conventional Gaussian policies may fail to fit the datasets with complex distributions for their *restricted expressiveness*. Meanwhile, although experience replay is used to improve sample efficiency [Mnih *et al.*, 2013], there is still *data scarcity* problem in environments with high-dimensional state spaces and complex interaction patterns. A common usage of learned dynamic models in model-based RL is planning in them [Nagabandi *et al.*, 2018; Schrittwieser *et al.*, 2020; Zhu *et al.*, 2021], but the per-step autoregressive planning approaches suffer from the *compounding error* problem [Xiao *et al.*, 2019]. An ideal RL algorithm should be able to learn a single policy to perform multiple tasks and generalize to new environments [Vithayathil Varghese and Mahmoud, 2020; Beck *et al.*, 2023]. However, existing works still struggle in *multitask generalizations*.

Recently, there has been a series of works applying diffusion models in sequential decision-making tasks, with a particular focus on offline RL. As a representative work, Diffuser [Janner *et al.*, 2022] fits a diffusion model for trajectory generation on the offline dataset, and plans desired future trajectories by guided sampling. There have been many following works where diffusion models behave as different modules in the RL pipeline, *e.g.*, replacing conventional Gaussian policies [Wang *et al.*, 2023], augmenting experience dataset [Lu *et al.*, 2023b], extracting latent skills [Venkataraman *et al.*, 2023], among others. We also observe that planning and decision-making algorithms facilitated by diffusion models perform well in broader applications such as multitask RL [He *et al.*, 2023a], imitation learning [Hegde *et al.*, 2023], and trajectory generation [Zhang *et al.*, 2022]. More importantly, diffusion models have already shed light on resolving those long-standing challenges in RL owing to their powerful and flexible distributional modeling ability.

¹<https://github.com/apexrl/Diff4RLSurvey>

This survey centers its attention on the utilization of diffusion models in RL, with additional consideration given to methods incorporating diffusion models in the contexts of trajectory generation and imitation learning, primarily due to the evident interrelations between these fields. Section 2 elaborates on the aforementioned RL challenges, and discusses how diffusion models can help solve each challenge. Section 3 provides a background on the foundations of diffusion models and also covers two class of methods that are particularly important in RL-related applications: guided sampling and fast sampling. Section 4 illustrates what roles diffusion models play in RL among existing works. Section 5 discusses the contribution of diffusion models on different RL-related applications. In Section 6, we point out the limitations when applying diffusion models and compare them with the transformer-based methods. Section 7 summarizes the survey with a discussion on emerging new topics.

2 Challenges in Reinforcement Learning

In this section, we list four challenges in RL algorithms and briefly discuss why diffusion models can address them.

2.1 Restricted Expressiveness in Offline Learning

Online RL [Sutton and Barto, 2018; Arulkumaran *et al.*, 2017] has been criticized for low sample efficiency, which makes it difficult to be applied in real-world scenarios. Offline RL [Fujimoto *et al.*, 2019; Kumar *et al.*, 2020; Fujimoto and Gu, 2021], which learns optimal policies purely from pre-collected datasets, obviates the need to interact with the environment during training, and can significantly improve sample efficiency. Directly applying off-policy RL methods [Lillicrap *et al.*, 2015; Haarnoja *et al.*, 2018] to offline learning suffers from the extrapolation error problem [Fujimoto *et al.*, 2019]. Existing works either penalize the value predictions on out-of-distribution samples [Kumar *et al.*, 2020; Nachum *et al.*, 2019], or constrain the learning policy to be close to the data collection policy [Wu *et al.*, 2019; Kostrikov *et al.*, 2021]. However, penalties on the value function can make the learned policy over-conservative [Lyu *et al.*, 2022]; for algorithms using policy constraints, since the policy is usually parameterized as a unimodal Gaussian, the restricted expressiveness makes it hard to fit the possibly diversified dataset. The reinforcement learning via supervised learning framework (RvS) [Schmidhuber, 2019; Srivastava *et al.*, 2019] is now becoming another essential paradigm in offline RL, which bypasses the need for Q learning and is thus free of extrapolation errors. RvS learns a policy conditioned on the observed returns via supervised learning and then conditions the learned policy on a high enough return during online evaluations to generate desired behaviors [Chen *et al.*, 2021; Lee *et al.*, 2022]. Similar to policy constraining methods, RvS requires fitting the entire offline dataset. Therefore, the expressiveness of parametrized policies also matters in RvS. Since diffusion models can represent arbitrary normalizable distributions [Neal and others, 2011], they hold potential to effectively improve the performance of policy constraining and RvS algorithms on complex datasets.

2.2 Data Scarcity in Experience Replay

Off-policy and offline RL methods use different levels of experience replay to improve sample efficiency. Note that experience replay in some cases only refers to data reuse in off-policy RL. Here, we use the term to broadly refer to updating the current model with rollout data from other policies. In off-policy RL, although all previously collected experiences can be used for policy learning, the available data during training may still be inadequate for effective policy optimization due to the speed limit of simulations and the potentially large state and action spaces. In offline RL, as no further interactions are allowed, policy learning is more limited by the quality and coverage of the dataset. Inspired by the success of data augmentation in computer vision, some works implement similar augmentation techniques in RL to mitigate the data scarcity problem. RAD [Laskin *et al.*, 2020] uses typical image augmentation techniques such as random cropping or rotation to boost the learning efficiency in vision-based RL. Imre [2021] and Cho *et al.* [2022] use generative models, VAE [Kingma and Welling, 2013] and GAN [Goodfellow *et al.*, 2014], to augment the real dataset with synthetic data sampled from the learned data distribution. However, existing works either suffer from a lack of fidelity when using random augmentation or are limited to simple environments due to insufficient modeling ability of particular generative models, making it difficult for them to extend to more complex tasks. Diffusion models have already demonstrated notable performances surpassing previous generative models in domains including image and video synthesis [Ho *et al.*, 2020; Ho *et al.*, 2022]. When applied to RL data, diffusion models are more suitable for augmenting high-dimensional datasets with intricate interactions.

2.3 Compounding Error in Model-based Planning

Model-based RL (MBRL) [Luo *et al.*, 2022; Moerland *et al.*, 2023] fits a model of dynamic transitions either from data obtained from online rollout or an offline dataset, and expects the model to facilitate decision-making. Common dynamic models mimic single-step state transitions and rewards in the dataset. Due to the limited data support and the possible stochasticity of the ground-truth transition function, there could be single-step errors when predicting with a neural network model. As a result, the compounding error problem arises when using the model for multistep planning [Xiao *et al.*, 2019], *i.e.*, cumulative single-step errors can cause the planned states to deviate from the dataset distribution, further increasing the error in the subsequent model predictions. In contrast, diffusion models with powerful modeling ability of joint distributions can operate on the trajectory level and plan for all time steps simultaneously, offering better temporal consistency and less compounding errors.

2.4 Generalization in Multitask Learning

Normal RL algorithms lack generalization abilities at the task level [Vithayathil Varghese and Mahmoud, 2020]. Even in the same environment, changing the reward function requires retraining a policy from scratch. Existing works studying online multitask RL [Yu *et al.*, 2020; Liu *et al.*, 2021] attempt to learn the same policy on different task environ-

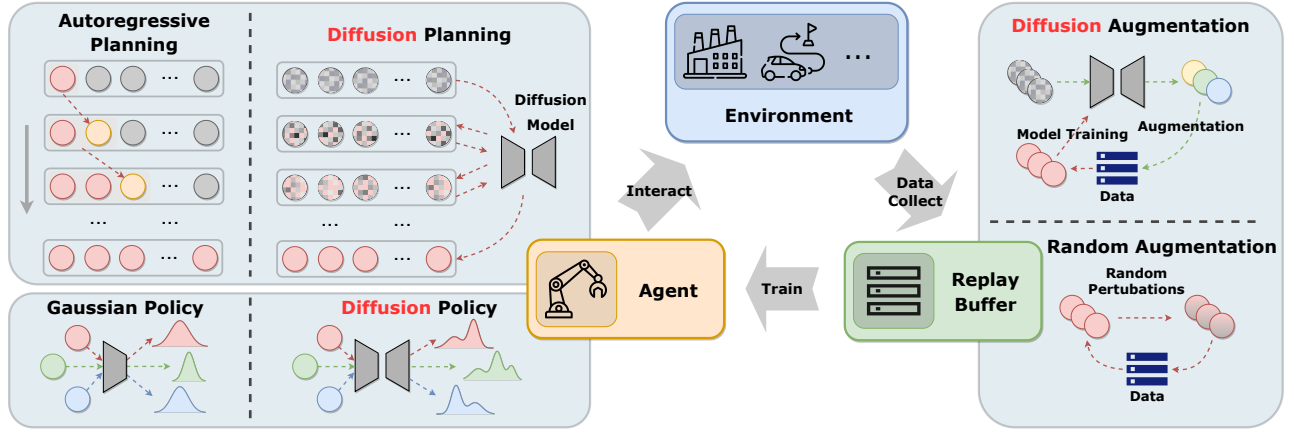


Figure 1: An illustration of how diffusion models play a different role in the classic Agent-Environment-Buffer cycle compared to previous solutions. (1) When used as a planner, diffusion models optimize the whole trajectory at each denoising step, whereas the autoregressive models generate the next-step output only based on previously planned partial subsequences. (2) When used as a policy, diffusion models can model arbitrary action distributions, whereas Gaussian policies can only fit the possibly diversified dataset distribution with unimodal distributions. (3) When used as a data synthesizer, diffusion models augment the dataset with generated data sampled from the learned dataset distribution, whereas augmentation with random perturbations might generate samples that deviate from data samples.

ments, suffering from the problem of conflicting gradients across multiple tasks, as well as low sample efficiency due to pure online learning. Recently, it has become a trending research direction to utilize a high-capacity model trained on multitask offline datasets and then deployed on new tasks with or without online fine-tuning [Taiga et al., 2022; Oh et al., 2017]. Transformer-based pre-training decision models like Gato [Reed et al., 2022] achieve notable successes in multitask policy learning. However, they typically require high-quality datasets and come with large parameter sizes as well as corresponding high training and inference costs. How to design an algorithm that can efficiently fit multitask datasets of mixed quality and generalize to new tasks emerges as a vital issue in multitask RL. As a powerful class of generative models, diffusion models can handle multimodal distributions in multitask datasets, and adapt to new tasks by estimating the task distribution.

3 Foundations of Diffusion Models

This section provides the foundations of diffusion models. Two prominent formulations are presented: Denoising Diffusion Probabilistic Model (DDPM) [Ho et al., 2020] and Score-based Generative Models [Song et al., 2021]. DDPM is widely used due to its simplicity, whereas the score-based formulation extends it to encompass continuous-time diffusion processes. Furthermore, guided sampling methods play a critical role in integrating diffusion models into RL frameworks. These methods can be divided into two main categories based on their approach to guiding the sampling process: classifier-guidance [Dhariwal and Nichol, 2021], which requires an extra classifier, and classifier-free guidance [Ho and Salimans, 2022], which makes guiding conditions as part of the model input. Additionally, in order to boost sampling speed especially during online interactions, fast sampling techniques have been adopted when us-

ing diffusion models in RL-related tasks [Kang et al., 2023; Chi et al., 2023]. Here we briefly cover some representative works in studying fast sampling of diffusion models, including learning-based and learning-free methods.

3.1 Denoising Diffusion Probabilistic Model

Assuming that the real data, denoted as x^0 , is sampled from an underlying distribution $q(x^0)$, the Denoising Diffusion Probabilistic Model (DDPM) utilizes a parameterized diffusion process, represented as $p_\theta(x^0) = \int p(x^T) \prod_{t=1}^T p_\theta(x^{t-1}|x^t) dx^{1:T}$, to model how the pure noise, denoted as $x^T = \mathcal{N}(0, I)$, is denoised into real data x^0 . In this formulation, each step of the diffusion process is represented by x^t , with T indicating the total number of steps. It is important to note that both diffusion models and RL use time step notations; thus, following common practice in RL, we denote diffusion steps with superscripts and RL time steps with subscripts. The sequence $x^{T:0}$ is defined as a Markov chain with learned Gaussian transitions characterized by

$$p_\theta(x^{t-1}|x^t) = \mathcal{N}(\mu_\theta(x^t, t), \Sigma(x^t, t)).$$

If the process is reversed to $x^{0:T}$, each step is the forward transition $q(x^t|x^{t-1})$. This transition can be interpreted as adding Gaussian noise to the data according to a variance schedule $\beta^{1:T}$:

$$x^t = \sqrt{\alpha^t} x^{t-1} + \sqrt{1 - \alpha^t} \epsilon_t, \quad (1)$$

where $\alpha^t = 1 - \beta^t$, $\epsilon_t \sim \mathcal{N}(0, I)$. q also holds the Markov property as $q(x^t|x^{t-1})$ only depends on x^{t-1} . In general, α^T should satisfy $\lim_{T \rightarrow +\infty} \alpha^T \rightarrow 0$ to ensure that x^T would be a pure standard Gaussian noise when T is sufficiently large. And from Eq. (1), we can infer that

$$x^t = \sqrt{\bar{\alpha}^t} x^0 + \sqrt{1 - \bar{\alpha}^t} \epsilon(x^t, t), \quad (2)$$

where $\bar{\alpha}^t = \prod_{i=1}^t \alpha^i$ and $\epsilon(x^t, t) \sim \mathcal{N}(0, \mathbf{I})$ is the integrated noise from step 0 to t . $\epsilon(x, t)$ is unknown and to be learned by a network. So far, we can calculate the distribution from step 0 to step t by simply one calculation. As for the denoising process $q(x^{t-1}|x^t)$, it is complex while $q(x^{t-1}|x^t, x^0)$ is relatively simpler. Based on Bayes Theorem and the Markov property of q , we have

$$q(x^{t-1}|x^0, x^t) = \frac{q(x^{t-1}|x^0)}{q(x^t|x^0)} q(x^t|x^{t-1}) = \mathcal{N}(\mu, \sigma^2),$$

where

$$\begin{aligned} \sigma^2(x^t, t) &= \beta^t \frac{1 - \bar{\alpha}^{t-1}}{1 - \bar{\alpha}^t}, \\ \mu(x^t, t) &= \frac{1 - \bar{\alpha}^{t-1}}{1 - \bar{\alpha}^t} \sqrt{\alpha^t} x^t + \frac{\sqrt{\bar{\alpha}^{t-1}} \beta^t}{1 - \bar{\alpha}^t} x^0. \end{aligned} \quad (3)$$

From Eq. (2), we can eliminate x^0 , yielding:

$$\mu(x^t, t) = \frac{1}{\sqrt{\alpha^t}} (x^t - \frac{\beta^t}{\sqrt{1 - \bar{\alpha}^t}} \epsilon(x^t, t)). \quad (4)$$

This allows us to sample x^T from standard Gaussian noise and progressively denoise it step by step until we obtain x^0 . However, the noise variable ϵ is still unknown. To address this, a network ϵ_θ , parameterized by θ , is employed to learn the noise generation process. In general, the loss function is the distance between the generated $\mu(x^t, t)$ using random variables ϵ and the network output ϵ_θ . Specially, $\mu(x^t, t)$ is defined as in Eq. (4), and $\mu_\theta(x^t, t) = \frac{1}{\sqrt{\alpha^t}} (x^t - \frac{\beta^t}{\sqrt{1 - \bar{\alpha}^t}} \epsilon_\theta(x^t, t))$.

Ho *et al.* [2020] discover that the following version of the loss function ignoring the weights has a better performance in experiments:

$$\mathcal{L}_{t-1} := \mathbb{E}_{x^0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\alpha^t} x^0 + (\sqrt{1 - \bar{\alpha}^t}) \epsilon, t)\|^2]. \quad (5)$$

Here, ϵ is sampled from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ and \mathcal{L}_{t-1} represents the loss function at diffusion step $t - 1$, when $t > 1$.

3.2 Score-based Generative Models

DDPM illustrates its iterative sampling process in a step-by-step manner, with each step taking place at a discrete time interval. Song *et al.* [2021] extend DDPM to continuous-time cases, where the sequence x^0, x^1, \dots, x^T is replaced with a continual function $x^t, t \in [0, T]$. Now, the forwarding process can be described as a Stochastic Differential Equation (SDE):

$$dx = f(x, t) dt + g(t) dw,$$

where $f(x, t)$ and $g(t)$ are pre-defined functions, and dw is the Brownian Motion. The example DDPM process in Eq. (1) is represented as $f(x, t) = -\frac{1}{2}\beta(t)x$, $g(t) = \sqrt{\beta(t)}$. According to Langevin Dynamics, the reverse of the forwarding process, as the sampling process, is described by a reverse-time SDE:

$$dx = [f(x, t) - g^2(t) \nabla_x \log q_t(x)] dt + g(t) d\bar{w}, \quad (6)$$

where \bar{w} is the reverse Brown Motion. The gradient term $\nabla_x \log q_t(x)$ is called the score function of the distribution

$q_t(x)$. As the score function remains unknown, score matching [Song *et al.*, 2019] proposes that a parameterized score-based model $\epsilon_\theta(x, t)$ can be trained to approximate it. In practice, diffusion models usually match the scaled score function $-\sigma_t \nabla_x \log q_t(x)$. Thus, a time-dependent score model can be trained by minimizing the Fisher Divergence:

$$\min_{\theta} \mathbb{E}_{x^0 \sim q_0, x^t \sim q_{0t}} [\|\epsilon_\theta(x^t, t) + \sigma_t \nabla_{x^t} \log q_{0t}(x^t|x^0)\|_2^2].$$

Solving the SDE in Eq. (6) is difficult even with a trained score model. Alternatively, we can solve the corresponding Probability Flow ODE (PF-ODE), whose marginal distribution at any time is equal to that given by the original SDE in Eq. (6):

$$dx = \left[f(x, t) - \frac{1}{2} g^2(t) \nabla_x \log q_t(x) \right] dt. \quad (7)$$

Given a trained score model, all terms in Eq. (7) are known thus it can be efficiently solved by various ODE solvers such as VODE [Brown *et al.*, 1989]. In this case, a black-box ODE solver [Dormand and Prince, 1980] is highly recommended because it not only produces high-quality samples but also allows us to explicitly trade-off accuracy for efficiency.

3.3 Guided Sampling Methods

Diffusion models with guided sampling methods care about the conditioned data distribution $p(x|y)$, which makes it possible to generate samples with attributes of the label y . Based on whether an extra classifier model is to be trained, the methods are divided into two categories: classifier guidance and classifier-free guidance. An advantage of classifier guidance sampling is that the classifier and the diffusion model are trained independently. If you already have a diffusion model, you just need to train a classifier and integrate it with your diffusion model while sampling. Classifier-free guidance re-trains the model totally, but it is more expressive and performs better compared to classifier guidance sampling methods.

Classifier Guidance

A simple thought is to train a differentiable discriminative model that presents $p(y|x)$. In other words, all we need to do is to train an extra classifier $p(y|x^t)$ trained on noisy samples x^t . Assume the classifier has been pre-trained. Then, the reversing process is described as

$$p_{\theta, \phi}(x^t|x^{t+1}, y) = Z p_{\theta}(x^t|x^{t+1}) p_{\phi}(y|x^t), \quad (8)$$

where Z is a normalization factor. Dhariwal and Nichol [2021] state that Eq. (8) can be approximately regarded as another Gaussian distribution:

$$p(x^t|x^{t+1}, y) = \mathcal{N}(\mu(x^t, t) + s \Sigma(x^t, t) g, \Sigma(x^t, t)),$$

where $g = \nabla_{x^t} \log p_{\phi}(y|x^t)|_{x^t=\mu}$ and s is the guidance scale to control the effect of conditions. The $\mu(x^t, t)$ and $\Sigma(x^t, t)$ function is identical to those in DDPM (Eq. (4) and Eq. (3)). Therefore, the sampling process can be explicitly presented as follows:

$$x^{t+1} \sim \mathcal{N}(\mu + s \Sigma g, \Sigma),$$

where Σ is short for $\Sigma(t, x^t)$.

Classifier-free Guidance

Rather than estimating the conditioned data distribution $p(x|y)$ directly, classifier-free methods try to predict the score function $\nabla_x \log p(x|y)$. With the Bayes Theorem, we can decompose the score function into an unconditional term and a classifier conditioning term as

$$\nabla_x \log p(x|y) = \nabla_x \log p(y|x) + \nabla_x \log p(x). \quad (9)$$

Unlike the classifier guidance, the original training setup is modified so the diffusion model should be retrained. The network to estimate noise is defined as $\epsilon(x^t, c)$ ($c = \emptyset$ for unconditional inputs). In practice, the conditional and unconditional models are trained with the same set of network parameters by randomly setting $c = \emptyset$ with a pre-specified probability during training. From Eq. (9) we can infer

$$\nabla_x \log p(y|x) = \nabla_x \log p(x|y) - \nabla_x \log p(x).$$

As Song *et al.* [2022] reveal, diffusion models and the score functions are equivalent, which indicates $\nabla_{x^t} \log p(x^t) \propto \epsilon(x^t, t)$. As a result, we can now substitute $\nabla_x \log p(y|x)$ with this into the formula for classifier guidance

$$\begin{aligned} \bar{\epsilon}_w(x^t, y) &= \epsilon_\theta(x^t, y) + w(\epsilon_\theta(x^t, y) - \epsilon_\theta(x^t)) \\ &= (1 + w)\epsilon_\theta(x^t, y) - w\epsilon_\theta(x^t), \end{aligned}$$

where w stands for the guidance scale.

3.4 Fast Sampling Methods

Diffusion model has been criticized for its prolonged iterative sampling time. Several fast sampling skills are proposed to solve this issue. Generally, the skills extend the diffusion model to a more fundamental paradigm to acquire efficiency. We categorize the methods into two categories: those that do not involve learning (learning-free) and those that require extra learning sessions (learning-based).

Learning-free sampling methods. Denoising Diffusion Implicit Models (DDIM) [Song *et al.*, 2022] is one of the earliest works on sampling acceleration. It aims at extending DDPM to a non-Markovian case by learning another Markov chain $q_\theta(x^{t-1}|x^t, x^0)$. The work also reveals that DDPM and DDIM are special cases of a more general paradigm. Moreover, it is observed in later works that DDIM is the discrete version of solving the PF-ODE Eq. (7). Then, some high-order solvers emerged, such as DPM-solver [Lu *et al.*, 2022], which provides an excellent trade-off between sample quality and sampling speed. With DDIM as its first-order version, DPM-solver boosts the efficiency of solving PF-ODE, outperforming common numerical ODE solvers like Runge-Kutta. Consequently, DPM-solver has become one of the most frequently used fast sampling methods.

Learning-based sampling methods. Learning-based sampling is another approach to fast sampling. Unlike the learning-free methods, they include extra learning processes to reach a higher sampling efficiency at a slight expense of sampling quality. A recent work, Truncate Diffusion Probabilistic Model (TDPM) [Zheng *et al.*, 2023], proposes that both the diffusion and denoising process can be truncated so that the iterative steps are reduced. Specifically, the forward

process is truncated when the sample is noisy enough, and the denoising process starts from a relatively noisy sample (not pure noise), which will be learned by a network or else. Moreover, Watson *et al.* [2021] learn a strategy to select the best K time steps to maximize the training objective for the DDPMs, which also decreases the denoising steps.

4 The Roles of Diffusion Models in RL

Diffusion models have proven their ability to generate diverse data and model multi-modal distributions. Considering the long-existing challenges introduced in Section 2, it is sufficient to improve the performance and sample efficiency of RL algorithms with diffusion models. In Fig. 1, we illustrate how diffusion models play a different role in RL compared to previous solutions. Current works applying diffusion models on RL mainly fall into three categories: using diffusion models as the planner, as the policy, and as the data synthesizer. It is essential to note that we include methods that generate action-only sequences as planners, even though some of the representative works have “policy” in their names, e.g., Diffusion Policy [Chi *et al.*, 2023]. Generating multi-step action sequences can be viewed as planning in action space, and the use of diffusion models to ensure temporal consistency is similar to other planning-based diffusion methods. The following subsections will illustrate overall frameworks and representative papers for each category.

4.1 Planner

Planning in RL refers to the process of using a model of the environment to make decisions imaginarily, and then selecting the best action in order to maximize a cumulative reward signal. This process usually simulates or explores different sequences of actions and states, predicting the outcomes of its decisions, thus resulting in better actions from the perspective of a longer horizon. Therefore, planning is commonly applied in the MBRL framework. However, the decision sequences used for planning are generated autoregressively, which may lead to severe compounding errors, especially in the offline setting, due to the limited data support. Diffusion models provide a possible solution since they can generate the whole sequence simultaneously.

A general framework of diffusion models as planners is shown in Fig. 2(a), which was first proposed by Janner *et al.* [2022]. Inputs and outputs of the diffusion model are usually clips of the real trajectory $\tau = (s_1, a_1, r_1, \dots, s_H, a_H, r_H)$, denoted as $x(\tau) = (e_1, e_2, \dots, e_H)$. Here e_t represents for the selected elements from (s_t, a_t, r_t) , where various choices can be made as $e_t = (s_t, a_t)$ [Janner *et al.*, 2022; Liang *et al.*, 2023; He *et al.*, 2023a; Xiao *et al.*, 2023], or $e_t = (s_t, a_t, r_t)$ [He *et al.*, 2023a; Hu *et al.*, 2023], $e_t = s_t$ [Ajay *et al.*, 2023; Zhu *et al.*, 2023], or $e_t = a_t$ [Chi *et al.*, 2023; Li *et al.*, 2023b]. $y(\tau)$ is the guidance that contains the desired property of the generated trajectory, such as the discounted return, the indicator of whether the task is completed, *etc.* Such guidance is meant to lead the diffusion model to generate good trajectories instead of those that only satisfy the environment dynamics. Suppose the diffusion model ϵ_θ parameterized by θ predicts the noise added in the forward

Table 1: Summary of papers on diffusion models for RL.

Model & Paper	Role of Diffusion Models	Keyword(s)	Guidance
Diffuser [Janner <i>et al.</i> , 2022]	Planner	Offline	Classifier
AdaptDiffuser [Liang <i>et al.</i> , 2023]		Offline	Classifier
EDGI [Brehmer <i>et al.</i> , 2023]		Offline	Classifier
TCD [Hu <i>et al.</i> , 2023]		Offline	Classifier-free
Crossway Diffusion [Li <i>et al.</i> , 2023b]		Offline; Robotics	None
SGP [Suh <i>et al.</i> , 2023]		Offline; Robotics	None
GSC [Mishra <i>et al.</i> , 2023]		Imitation; Robotics	Classifier
UniPi [Du <i>et al.</i> , 2023a]		Imitation; Robotics	Classifier-free
ChainedDiffuser [Xian <i>et al.</i> , 2023]		Imitation; Robotics	None
Diffusion Policy [Chi <i>et al.</i> , 2023]		Imitation; Robotics	None
AVDC [Ko <i>et al.</i> , 2023]		Imitation; Robotics	None
MTDiff-p [He <i>et al.</i> , 2023a]		Offline; Multi-task	Classifier-free
MetaDiffuser [Ni <i>et al.</i> , 2023]		Offline; Multi-task	Classifier-free
SafeDiffuser [Xiao <i>et al.</i> , 2023]		Offline; Safe	Self-proposed
MADiff [Zhu <i>et al.</i> , 2023]		Offline; Multi-agent	Classifier-free
HDMI [Li <i>et al.</i> , 2023a]		Offline; Hierarchical	Classifier-free
MLD [Chen <i>et al.</i> , 2022]		Trajectory Generation	Classifier-free
MDM [Tevet <i>et al.</i> , 2022]		Trajectory Generation	Classifier-free
UniSim [Yang <i>et al.</i> , 2023]		Trajectory Generation	Classifier-free
ReMoDiffuse [Zhang <i>et al.</i> , 2023b]		Trajectory Generation	Classifier-free
SinMDM [Raab <i>et al.</i> , 2023]		Trajectory Generation	None
EquiDiff [Chen <i>et al.</i> , 2023b]		Trajectory Generation	None
MoFusion [Dabral <i>et al.</i> , 2022]		Trajectory Generation	None
MotionDiffuse [Zhang <i>et al.</i> , 2022]		Trajectory Generation	None
MPD [Carvalho <i>et al.</i> , 2023]	Policy	Trajectory Generation; Robotics	Classifier
MotionDiffuser [Jiang <i>et al.</i> , 2023]		Trajectory Generation; Multi-agent	Classifier
AlignDiff [Dong <i>et al.</i> , 2023]		RLHF	Classifier-free
Diffusion-QL [Wang <i>et al.</i> , 2023]		Offline	Q-loss
SRDP [Ada <i>et al.</i> , 2023]		Offline	Q-loss
SfBC [Chen <i>et al.</i> , 2023a]		Offline	Policy Gradient
IDQL [Hansen-Estruch <i>et al.</i> , 2023]		Offline	Policy Gradient
EDP [Kang <i>et al.</i> , 2023]		Offline	Policy Gradient
DiffCPS [He <i>et al.</i> , 2023b]		Offline	None
NoMaD [Sridhar <i>et al.</i> , 2023]		Robotics	None
BESO [Reuss <i>et al.</i> , 2023]		Offline; Goal-conditioned	Classifier-free
CEP [Lu <i>et al.</i> , 2023a]		Offline; Image Synthesis	Policy Gradient
DOM2 [Li <i>et al.</i> , 2023c]		Offline; Multi-agent	Q-loss
CPQL [Chen <i>et al.</i> , 2023d]		Offline; Online	Q-loss
Pearce <i>et al.</i> [2023]		Imitation	Classifier-free
Yoneda <i>et al.</i> [2023]		Imitation; Robotics	None
PlayFusion [Chen <i>et al.</i> , 2023c]		Imitation; Robotics	None
XSkill [Xu <i>et al.</i> , 2023]		Imitation; Robotics	None
CoDP [Ng <i>et al.</i> , 2023]		Human-in-the-loop	None
GenAug [Chen <i>et al.</i> , 2023e]	Data Synthesizer	Robotics	None
ROSIE [Yu <i>et al.</i> , 2023]		Robotics	None
SynthER [Lu <i>et al.</i> , 2023b]		Offline; Online	None
MTDiff-s [He <i>et al.</i> , 2023a]		Offline; Multi-task	Classifier-free
LDCQ [Venkatraman <i>et al.</i> , 2023]	Latent Representation	Offline	Classifier-free
LatentDiffuser [Li, 2023]		Offline	Policy Gradient
Hegde <i>et al.</i> [2023]		Quality Diversity	None
DVF [Mazouze <i>et al.</i> , 2023]	Value Function	Offline	None

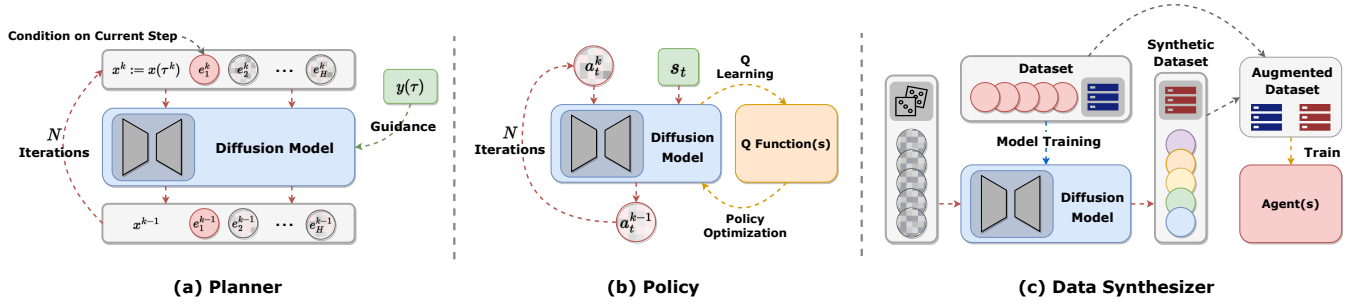


Figure 2: Different roles of diffusion models in RL. (a) Diffusion models as the planner. The sampling target is a part of trajectories whose components may vary from specific tasks. (b) Diffusion models as the policy. The sampling target is the action conditioned on the state, usually guided by the Q-function via policy gradient-style guidance or directly subtracting it from the training objective. (c) Diffusion models as the data synthesizer. The sampling target is also the trajectory, and both real and synthetic data are used for downstream policy improvement. For better visualizations, we omit the arrows for N denoising iterations in (c) and only show generated synthetic data from randomly sampled noise. Note that there are other roles that are less explored, and we introduce them in Section 4.4.

process, and $x^k := x(\tau^k)$ is the corresponding clip of τ at the k -th diffusion step. Either classifier guidance or classifier-free guidance is incorporated. For classifier guidance [Janner et al., 2022; Liang et al., 2023], ϵ_θ is trained as usual by taking τ^k and the timestep k as input:

$$\mathcal{L}(\theta) = \mathbb{E}_{\tau, k, \epsilon} [\|\epsilon - \epsilon_\theta(x^k, k)\|_2^2].$$

After training, suppose $\mu_\theta(x^k, k)$ and Σ are the mean and the variance of the original reverse process $p_\theta(x^{k-1}|x^k)$, respectively. The gradient of $y(\tau)$ is injected during sampling as the guidance:

$$p_\theta(x^{k-1}|x^k, y(\tau)) \approx \mathcal{N}(x^{k-1}; \mu_\theta + \alpha \Sigma^k \nabla y(\tau), \Sigma^k),$$

where α is a hyperparameter. In contrast, classifier-free guidance [Ajay et al., 2023; He et al., 2023a; Hu et al., 2023] injects $y(\tau)$ in both training and sampling stages. The noise prediction model ϵ_θ now takes an additional $y(\tau)$ as input, giving the training objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\tau, k, \epsilon, \beta} [\|\epsilon - \epsilon_\theta(x^k, (1 - \beta)y(\tau) + \beta\emptyset, k)\|_2^2],$$

where β is sampled from a Bernoulli distribution with probability p . At sampling, $y(\tau)$ is directly fed into ϵ_θ , as

$$p_\theta(x^{k-1}|x^k, y(\tau)) \approx \mathcal{N}(x^{k-1}; \mu_\theta(x^k, y(\tau), k), \Sigma^k).$$

As the time consumption of sampling using the original DDPM is unacceptable when deploying RL in the online environment, fast sampling methods are commonly incorporated, and the diffusion step for sampling is usually set to 100 or 200. Besides, to guarantee sample stability during the sampling process, the first several steps of a sampled trajectory can be replaced with the ground truth as a hard constraint. For example, when planning the trajectory after $e_{1:h}$, the model first generates the full trajectory $\tilde{x}^k = \tilde{e}_{1:H}^k$ at the k -th diffusion step. Then the first h steps are replaced, forming $\hat{x}^k \leftarrow (e_1, e_2, \dots, e_h, \tilde{e}_{h+1}^k, \dots, \tilde{e}_H^k)$.

The core ideas behind why non-autoregressive diffusion models can generate Markovian decision sequences are stated by Janner et al. [2022]. With the U-Net backbone, the

sampled token e_t^k takes information in a local receptive field $e_{t-l:t+l}^{k+1}$, where l is determined by the model structure. Once the local consistency is guaranteed, the global consistency will also be built during the denoising process as information spreads among the trajectory. Though this raises new confusion that using future information is anti-causal for sequential problems, RL includes a hidden hypothesis that the current decision will lead to optimal future outcomes. That is, utilizing future information does not violate the general RL framework.

Benefiting from the non-autoregressive scheme, diffusion models surpass traditional models in generating planning sequences. First, such generations do not suffer from the aforementioned compounding errors. Then, diffusion models fit for both continuous-reward and sparse-reward settings [Janner et al., 2022], even for more challenging multitask planning [He et al., 2023a], safe planning [Xiao et al., 2023], and generating based on multi-modal observations [Du et al., 2023a; Chi et al., 2023]. Besides, diffusion models can also learn from non-Markovian data by exploiting temporal information [Hu et al., 2023].

4.2 Policy

Compared with traditional RL taxonomy, which roughly divides RL algorithms into MBRL and model-free RL, using diffusion models as the planner is similar to MBRL and focuses on capturing the environment dynamics. In contrast, taking diffusion models as the policy follows the framework of model-free RL. Section 2.1 states the main drawbacks of offline policy learning frameworks: over-conservatism and poor capability on diversified datasets. With excellent expressive ability on multi-modal distribution, many works utilize diffusion models as the policy to tackle these problems.

Diffusion-QL [Wang et al., 2023] first explores the advantages of the diffusion policy in offline RL and finds that it can perfectly fit on dataset generated by strong multi-modal behavior policies, where previous distance-based BC methods fail. Experiment results also support the dominance of diffusion policies on diversified datasets [Lu et al., 2023a; Chen

et al., 2023a; Hansen-Estruch *et al.*, 2023; Ada *et al.*, 2023; Li *et al.*, 2023c]. Compared with diffusion models as planners, the diffusion target of the diffusion policy is simply the action given the current state, as shown in Fig. 2(b). Suppose the noise predictor is $\epsilon_\theta(a^k, k, s)$ parameterized by θ , and the derived diffusion model is $\mu_\theta(a|s)$. To guide the model sampling actions with high Q-values, it is necessary to take $Q(s, a)$ into consideration. Some papers [Chen *et al.*, 2023a; Lu *et al.*, 2023a; Hansen-Estruch *et al.*, 2023; Kang *et al.*, 2023] construct the policy by (advantage) weighted regression as

$$\pi_\theta(a|s) \propto \mu_\theta(a|s) \exp(\alpha Q(s, a)),$$

where α is a hyperparameter. Following this, Chen *et al.* [2023a] decouple the policy learning into behavior learning and action evaluation, which allows more freedom in the choice of guidance. It also proposes in-sample planning for Q-learning, which avoids extrapolation errors in previous offline RL methods. Lu *et al.* [2023a] further generalize this framework to any energy reweighted distribution $p(x) \propto q(x) \exp(-\beta \mathcal{E}(x))$, where $\mathcal{E}(x)$ is an energy function, and train it via contrastive learning. Approaches using explicit Q-values as guidance are also considered. Wang *et al.* [2023] and Ada *et al.* [2023] subtract a weighted expectation of $Q(s, a)$ in the training loss as

$$\begin{aligned} \mathcal{L}(\theta) = & \mathbb{E}_{k, \epsilon, (s, a) \sim \mathcal{D}} [\|\epsilon - \epsilon_\theta(a^k, s, k)\|_2^2] \\ & - \frac{\eta}{\mathbb{E}_{(s, a) \sim \mathcal{D}} [Q(s, a)]} \cdot \mathbb{E}_{s \sim \mathcal{D}, a^0 \sim \pi_\theta(\cdot|s)} [Q(s, a^0)], \end{aligned}$$

where η is a hyperparameter, and \mathcal{D} is the offline dataset.

Fast reaction is crucial when deploying policies in online environments. Therefore, almost all diffusion policies use smaller diffusion steps during sampling, usually about 15 steps. ODE solvers such as the DPM-solver [Lu *et al.*, 2022] are also used to accelerate sampling [Chen *et al.*, 2023a; Lu *et al.*, 2023a; Kang *et al.*, 2023; Li *et al.*, 2023c]. Pearce *et al.* [2023] compare three different backbone structures of diffusion models, finding that MLP-sieve or Transformer is also sufficient in expressiveness. Kang *et al.* [2023] introduce action approximation, which allows one-step action sampling in the training stage.

4.3 Data Synthesizer

In addition to fitting multi-modal distributions, a simple and common use of diffusion models is to generate more training samples, which has been widely applied and proven in computer vision. Therefore, it is natural to apply the diffusion model as a data synthesizer on RL datasets, as data scarcity is a practical challenge of RL, as stated in Section 2.2. To guarantee consistency of synthetic data to the environment dynamics, previous data augmentation approaches in RL usually add small perturbations to existing states and actions [Sinha *et al.*, 2021]. In contrast, Fig. 2(c) illustrates that diffusion models learn the data distribution from the whole dataset \mathcal{D} , and enable generating highly diversified data while keeping consistency. Lu *et al.* [2023b] investigate the ability of diffusion models as the data synthesizer in both offline and online settings. It directly trains the diffusion model from the offline dataset or the online replay buffer and then generates

more samples for policy improvement. Analysis shows that the quality of data generated by diffusion models is higher than those generated by explicit data augmentation in diversity and accuracy. With synthetic data, the performance of the offline policy and the sample efficiency of the online policy are significantly improved. He *et al.* [2023a] deploy diffusion models to augment data for multitask offline datasets and achieves better performance than those on single-task datasets. It claims that fitting on multiple tasks may enable implicit knowledge sharing across tasks, which also benefits from the multi-modal property of diffusion models.

4.4 Others

Besides directions that have been mainly focused on, some ways of improving RL with diffusion models are under exploration. Hegde *et al.* [2023] take a similar idea as hyper networks in meta-learning, generating parameters of policies for quality diversity RL. The trained diffusion models compress the parameters of various policies into the latent space while maintaining the ability to generate policies under certain conditions. Mazouze *et al.* [2023] estimate value functions with diffusion models by learning the discounted state occupancy, combined with a learned reward estimator. Then, the value function can be directly computed by definition, where future states are sampled from the diffusion model. Venkatraman *et al.* [2023] follow Latent Diffusion Models [Rombach *et al.*, 2022] by first encoding the high-level trajectories into semantically rich representations, then applying diffusion models on them. Conditioning on latent representations, Q-functions, and policies achieves higher capability without significant extrapolation errors.

5 Applications of Diffusion Models

Diffusion models have recently been employed to address various problems in RL and several strongly related fields. We group these applications into four categories based on the task: offline RL, imitation learning, trajectory generation, and data augmentation. For each category, we provide a brief introduction to the task, followed by a detailed explanation of how existing works use diffusion models to improve performance on that task.

5.1 Offline RL

Offline RL [Levine *et al.*, 2020] aims to learn a policy from previously collected datasets without online interaction. Assuming there is a static dataset \mathcal{D} collected by some (unknown) behavior policy π_β , offline RL requires the learning algorithm to derive a policy $\pi(a|s)$ that attains the most cumulative reward, which is defined in Eq. (10). The fundamental challenge in offline RL is the distributional shift: while the function approximator (*e.g.*, policy, value function) might be trained under one distribution, it will be evaluated on a different distribution, leading to poor performance of the learned policy. High-dimensional and expressive function approximation generally exacerbates this issue:

$$\pi^* := \arg \max_{\pi} \mathbb{E}_{\tau \sim p_{\pi}(\tau)} \left[\sum_{t=0}^H \gamma^t r(s_t, a_t) \right]. \quad (10)$$

Several methods use diffusion models to help tackle or avoid the above challenge. [Janner et al. \[2022\]](#) first propose to generate optimal trajectories through iteratively denoising with classifier-guided sampling. Subsequent works [[Wang et al., 2023](#); [Chen et al., 2023a](#); [He et al., 2023b](#); [Ada et al., 2023](#); [Brehmer et al., 2023](#); [Hansen-Estruch et al., 2023](#); [Venkatraman et al., 2023](#)] represent the policy as a diffusion model to capture multi-modal distributions and enhance the expressiveness of the policy class, which is beneficial to relieve the approximation error between the cloned behavior policy and true behavior policy. [Ajay et al. \[2023\]](#) skip the risk of distribution shift by generating state sequences with conditional diffusion models followed by inverse dynamic functions to derive executable actions, which propose a novel approach to use classifier-free guidance with low-temperature sampling to denoise out return-maximizing trajectories. In order to improve the generation ability of diffusion models for RL, [Lu et al. \[2023a\]](#) propose a new guidance method named contrastive energy prediction and [Hu et al. \[2023\]](#) capture more temporal conditions. By incorporating control-theoretic invariance into the diffusion dynamics, SafeDiffuser [[Xiao et al., 2023](#)] guarantees the safe generation of planning trajectories. HDMI [[Li et al., 2023a](#)] leverages a hierarchical structure to tackle long-horizon decision-making problems, which uses a reward-conditional model to discover sub-goals and a goal-conditional model to generate actions. AlignDiff [[Dong et al., 2023](#)] conditions on behavior attributes with classifier-free guidance to plan to match desired trajectories accurately. CPQL [[Chen et al., 2023d](#)] leverages consistency models for fast training and sampling, while EDP [[Kang et al., 2023](#)] obtains speed-up by using single-step model predictions as action approximations. Diffusion models can also be used to extract the reward function [[Nuti et al., 2023](#)] or value function [[Mazouze et al., 2023](#)], and are prominent to estimate gradients with score-matching to solve offline optimization problems [[Suh et al., 2023](#)]. Enabling RL agents to generalize to multi-task and multi-agent scenarios remains a challenge due to their inherent complexity. Recent research has made progress in using diffusion models to improve the performance of policies in multi-task and multi-agent offline RL.

Multitask offline RL. Diffusion model is verified to have the potential to address the challenge of multi-task generalization in RL. [He et al. \[2023a\]](#) first extend the conditional diffusion model to be capable of solving multitask decision-making problems and synthesizing useful data for downstream tasks. LCD [[Zhang et al., 2023a](#)] leverages a hierarchical structure to achieve long-horizon multi-task control. [Ni et al. \[2023\]](#) and [Liang et al. \[2023\]](#) extend the idea of Diffuser [[Janner et al., 2022](#)] into more specific settings. MetaDiffuser [[Ni et al., 2023](#)] demonstrates that incorporating the conditional diffusion model into the context of task inference outperforms previous meta-RL methods. AdaptDiffuser [[Liang et al., 2023](#)] combines bootstrapping and diffusion-based generative modeling together to enable the model to adapt to unseen tasks.

Multi-agent offline RL. Employing diffusion models to multi-agent RL helps model discrepant behaviors among agents and reduces approximation error. MADiff [[Zhu et al.,](#)

[2023](#)] uses an attention-based diffusion model to model the complex coordination among behaviors of multiple agents, which is well-suited to learning complex multi-agent interactions. DOM2 [[Li et al., 2023c](#)] incorporates the diffusion model into the policy classes to enhance learning and makes it possible to generalize to shifted environments well.

5.2 Imitation Learning

The goal of imitation learning (IL) is to reproduce behavior similar to experts in the environment by extracting knowledge from expert demonstrations. Recently, many works [[Hegde et al., 2023](#); [Ng et al., 2023](#); [Chen et al., 2023c](#); [Kapelyukh et al., 2022](#)] have demonstrated the efficacy of representing policies as diffusion models to capture multi-modal behavior. [Pearce et al. \[2023\]](#) apply diffusion models to imitate human behavior in sequential environments, in which diffusion models are compared with other generative models and viable approaches are developed to improve the quality of behavior sampled from diffusion models. [Chi et al.; Xian et al. \[2023; 2023\]](#) generate the robot’s behavior via a conditional denoising diffusion process on robot action space. Experiment results show that Diffusion models are good at predicting closed-loop action sequences while guaranteeing temporal consistency [[Chi et al., 2023](#)]. [Li et al. \[2023b\]](#) improve the models in [Chi et al. \[2023\]](#) by incorporating an auxiliary reconstruction loss on intermediate representations of the reverse diffusion process. Beneficial from its powerful generation ability, leveraging diffusion models to acquire diverse skills to handle multiple manipulation tasks is promising [[Chen et al., 2023c](#); [Mishra et al., 2023](#); [Xu et al., 2023](#); [Ha et al., 2023](#)]. Diffusion models are already applied to goal-conditioned RL: [Reuss et al. \[2023\]](#) use a decoupled score-based diffusion model to learn an expressive goal-conditional policy. In contrast, [Sridhar et al. \[2023\]](#) build a unified diffusion policy to solve both goal-directed navigation and goal-agnostic exploration problems.

5.3 Trajectory Generation

Trajectory generation aims to derive a dynamically feasible path that satisfies a set of constraints. In particular, we focus on generating human pose and robot interaction sequences, which are more related to the decision-making scenario. Many works [[Zhang et al., 2022](#); [Jiang et al., 2023](#); [Tevet et al., 2022](#); [Zhang et al., 2023b](#); [Chen et al., 2022](#); [Dabral et al., 2022](#)] have remarked that the conditional diffusion models perform better than traditional methods which use GAN or Transformer. Employing a denoising-diffusion-based framework, they achieve diverse and fine-grained motion generation with various conditioning contexts [[Chen et al., 2023b](#); [Carvalho et al., 2023](#)]. Recent works [[Du et al., 2023b](#); [Ko et al., 2023](#); [Du et al., 2023a](#)] harness diffusion models to synthesize a set of future frames depicting its planned actions in the future, after which control actions are extracted from the generated video. This approach makes it possible to train policies solely based on RGB videos and deploy learned policies to various robotic tasks [[Black et al., 2023](#); [Gao et al., 2023](#)]. UniSim [[Yang et al., 2023](#)] uses diffusion models to build a universal simulator of real-world interaction by learning through combined diverse datasets. It can

be utilized to train both high-level vision-language planners and low-level RL policies, demonstrating powerful emulation ability.

5.4 Data Augmentation

Diffusion models have already been verified to be useful for data augmentation in the RL domain. Since diffusion models perform well in learning over multimodal or even noisy distributions, they can model original data distribution precisely. What is more, they are capable of generating diverse data points to expand original distribution while maintaining dynamic accuracy. Recent works [Yu *et al.*, 2023; Chen *et al.*, 2023e] consider augmenting the observations of robotic control using a text-guided diffusion model while maintaining the same action. The recently proposed SynthER [Lu *et al.*, 2023b] and MTDiff-s [He *et al.*, 2023a] generate complete transitions of trained tasks via a diffusion model. It proves that such augmentation brings about significant policy improvement for both online and offline RL.

6 Limitations and Remarks

In this section, we list three limitations when applying diffusion models in RL, and include remarks on comparing diffusion-based generative modeling in RL to transformer-based autoregressive approaches.

Application in online RL. Though diffusion models have contributed to offline RL from various perspectives, the dynamic and evolving nature of online RL introduces a significant challenge, as the data distribution sampled by the current policy can change over time. The primary concern is that effectively adapting diffusion models to changing data distributions requires a substantial amount of new data [Lu *et al.*, 2023b]. These models are typically trained on fixed datasets, and incorporating enough data to ensure they can generalize and remain effective in dynamic real-world scenarios is a resource-intensive task. Balancing the need for adaptability with the requirement for extensive data is the primary consideration when applying diffusion models in online RL. It is promising to solve this dilemma with more lightweight diffusion models which can keep consistency as the data distribution changes during online interactions.

Iterative sampling cost. The unique stepwise denoising mechanism of diffusion models makes the sampling procedure need to infer the trained model multiple times, significantly increasing the sampling cost. Although we can perform sampling acceleration techniques such as DDIM or DPM-Solver, the expensive inference cost limits the model to conduct high-frequency output in online interactions with virtual environments and real-world continuous control scenarios. The problem becomes more severe in those methods that generate a long trajectory but only take the first state or action for execution, such as Diffuser [Janner *et al.*, 2022] or Decision Diffuser [Ajay *et al.*, 2023]. Chen *et al.* [2023d] incorporate the recently proposed Consistency Model [Song *et al.*, 2023] to enable sampling with one or two diffusion steps, and the performances are comparable to that of DDPM or DDIM with 50 steps.

Variance in stochastic sampling. In traditional RL algorithms, a continuous control policy is represented by a state-conditioned Gaussian [Haarnoja *et al.*, 2018; Schulman *et al.*, 2017]. We can take its mean as action when deterministic execution is required. However, such deterministic policy is not possible when using diffusion models as the policy class. The randomness of diffusion sampling comes from both initial noise and per-step stochastic denoising. Although per-step stochasticity can be avoided if sampling with ODE-based methods like DDIM, the randomness in initial noisy samples remains inevitable. The high-variance policies can have a negative impact in environments with high accuracy or safety requirements. Existing works in RL rarely discuss this limitation, and sampling methods with reduced variances are expected.

Comparison to transformer-based methods. The idea of using diffusion models commonly abstracts offline RL as a conditional generative modeling problem, differing from traditional RL approaches, which need value function approximations or policy gradient calculation. This supervised learning paradigm is similar to Decision Transformer (DT) [Chen *et al.*, 2021], a sequential modeling framework for RL tasks with transformer architecture. The essence of both approaches is to take advantage of the high expressiveness model, where diffusion models enjoy a strong distribution fitting ability to produce multi-modal, diverse, and accurate outputs, and transformers are adept in long-horizon sequence modeling and time correlation understanding. Qualitatively, this difference in expertise makes the diffusion-based approach more suitable for learning complex multi-modal tasks, and Transformer-based approaches [Meng *et al.*, 2021; Wen *et al.*, 2022] are more preferred in correlated sequence modeling in time or agent (in multi-agent tasks) dimension.

7 Summary and Future Prospects

This survey offers a comprehensive overview of contemporary research endeavors concerning the application of diffusion models in the realm of RL. According to the roles played by diffusion models, we categorize existing methodologies into using diffusion models as planners, policies, data synthesizers, and less popular roles such as value functions, latent representation models, *etc.* By comparing each class of methods to traditional solutions, we can see how the diffusion model addresses some of the longstanding challenges in RL, *i.e.*, restricted expressiveness, data scarcity, compounding error, and multitask generalization. Notwithstanding these merits, it is imperative to acknowledge the existence of non-negligible limitations in using diffusion models in RL due to some inherent properties in the training and sampling of diffusion models. It is worth emphasizing that the incorporation of diffusion models into RL remains an emerging field, and there are many research topics worth exploring. Here, we outline three prospective research directions, namely, retrieval-enhanced generation, integrating safety constraints, and composing different skills.

Retrieval-enhanced generation. Retrieval techniques are employed in various domains such as recommender systems [Qin *et al.*, 2020] and large language models [Kandpal

et al., 2023] to enhance the model capacity and handle long-tail distributed datasets. Some works utilize retrieved data to boost text-to-image and text-to-motion diffusion generation [Sheynin *et al.*, 2022; Zhang *et al.*, 2023b], promoting better coverage of uncommon condition signals. During on-line interactions, RL agents may also encounter states that are rare in the training dataset. By retrieving relevant states as model inputs, the performance of diffusion-based decision models can be improved in these states. Also, if the retrieval dataset is constantly updated, diffusion models have the potential to generate new behaviors without retraining.

Integrating safety constraints. Utilizing RL models for practical applications often necessitates compliance with various safety constraints [García and Fernández, 2015; Gu *et al.*, 2022]. Several safe RL methods transform a constrained RL problem to its unconstrained equivalent [Achiam *et al.*, 2017; Tessler *et al.*, 2018; Sootla *et al.*, 2022], which is then solved by generic RL algorithms. Policies acquired through these methods remain tailored to the specific constraint threshold utilized during training. Recent research [Liu *et al.*, 2023; Zhang *et al.*, 2023c] has extended the applicability of decision transformers to the context of safety-constrained settings, thereby enabling a single model to adapt to diverse thresholds by adjusting the input cost-to-go. Similarly, diffusion models exhibit substantial potential for deployment in the domain of safe RL. Ajay *et al.* [2023] demonstrate that a diffusion-based planner can combine different movement skills to produce new behaviors. In addition, classifier-guided sampling can add new conditions to generated samples simply by learning additional classifiers, while the parameters of the diffusion model itself remain unchanged [Dhariwal and Nichol, 2021]. This feature makes the diffusion model promising to be effective in scenarios where new safety constraints can be included after model training.

Composing different skills. Most current works deploy the generation ability of diffusion models on the raw state and action spaces. From the perspective of skill-based RL [Shi *et al.*, 2022; Nam *et al.*, 2022], it is promising to break down complex tasks into smaller, more manageable sub-skills. Diffusion models excel in modeling multi-modal distributions, and since multiple sub-skills can be viewed as distinct modes within the distribution of possible behaviors, they offer a natural fit for this task. Combining with classifier guidance or classifier-free guidance, diffusion models are possible to generate proper skills to complete the facing task. Experiments in offline RL also suggest that diffusion models can share knowledge across skills and combine them up [Ajay *et al.*, 2023; He *et al.*, 2023a], thus having the potential for zero-shot adaptation or continuous RL by composing different skills.

Acknowledgments

The work is partially supported by National Key R&D Program of China (2022ZD0114804) and National Natural Science Foundation of China (62076161). We thank Minghuan Liu, Xihuai Wang, Jingxiao Chen and Mingcheng Chen for valuable suggestions and discussions.

References

- [Achiam *et al.*, 2017] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.
- [Ada *et al.*, 2023] Suzan Ece Ada, Erhan Oztop, and Emre Ugur. Diffusion policies for out-of-distribution generalization in offline reinforcement learning, 2023.
- [Ajay *et al.*, 2023] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, T. Jaakkola, and Pulkrit Agrawal. Is conditional generative modeling all you need for decision-making? 2023.
- [Arulkumaran *et al.*, 2017] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.
- [Austin *et al.*, 2021] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.
- [Beck *et al.*, 2023] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.
- [Black *et al.*, 2023] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models, 2023.
- [Brehmer *et al.*, 2023] Johann Brehmer, Joey Bose, Pim de Haan, and Taco Cohen. Edgi: Equivariant diffusion for planning with embodied agents, 2023.
- [Brown *et al.*, 1989] Peter N. Brown, George D. Byrne, and Alan C. Hindmarsh. Vode: A variable-coefficient ode solver. *SIAM Journal on Scientific and Statistical Computing*, 10(5):1038–1051, 1989.
- [Carvalho *et al.*, 2023] Joao Carvalho, An T. Le, Mark Baierl, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and planning of robot motions with diffusion models, 2023.
- [Chen *et al.*, 2021] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.
- [Chen *et al.*, 2022] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2022.
- [Chen *et al.*, 2023a] Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, and Jun Zhu. Offline reinforcement learning via high-fidelity generative behavior modeling. In *The*

Eleventh International Conference on Learning Representations, 2023.

- [Chen *et al.*, 2023b] Kehua Chen, Xianda Chen, Zihan Yu, Meixin Zhu, and Hai Yang. Equidiff: A conditional equivariant diffusion model for trajectory prediction, 2023.
- [Chen *et al.*, 2023c] Lili Chen, Shikhar Bahl, and Deepak Pathak. Playfusion: Skill acquisition via diffusion from language-annotated play. In *7th Annual Conference on Robot Learning*, 2023.
- [Chen *et al.*, 2023d] Yuhui Chen, Haoran Li, and Dongbin Zhao. Boosting continuous control with consistency policy, 2023.
- [Chen *et al.*, 2023e] Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *arXiv preprint arXiv:2302.06671*, 2023.
- [Chi *et al.*, 2023] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2023.
- [Cho *et al.*, 2022] Daesol Cho, Dongseok Shim, and H Jin Kim. S2p: State-conditioned image synthesis for data augmentation in offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:11534–11546, 2022.
- [Dabral *et al.*, 2022] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Motion: A framework for denoising-diffusion-based motion synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9760–9770, 2022.
- [Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [Dong *et al.*, 2023] Zibin Dong, Yifu Yuan, Jianye Hao, Fei Ni, Yao Mu, Yan Zheng, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model, 2023.
- [Dormand and Prince, 1980] J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980.
- [Du *et al.*, 2023a] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023.
- [Du *et al.*, 2023b] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning, 2023.
- [Fujimoto and Gu, 2021] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.
- [Fujimoto *et al.*, 2019] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.
- [Gao *et al.*, 2023] Jialu Gao, Kaizhe Hu, Guowei Xu, and Huazhe Xu. Can pre-trained text-to-image models generate visual goals for reinforcement learning?, 2023.
- [García and Fernández, 2015] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.
- [Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [Gu *et al.*, 2022] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.
- [Ha *et al.*, 2023] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *7th Annual Conference on Robot Learning*, 2023.
- [Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.
- [Hansen-Estruch *et al.*, 2023] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies, 2023.
- [He *et al.*, 2023a] Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. 2023.
- [He *et al.*, 2023b] Longxiang He, Linrui Zhang, Junbo Tan, and Xueqian Wang. Diffcps: Diffusion model based constrained policy search for offline reinforcement learning, 2023.
- [Hegde *et al.*, 2023] Shashank Hegde, Sumeet Batra, K. R. Zentner, and Gaurav S. Sukhatme. Generating behaviorally diverse policies with latent diffusion models, 2023.
- [Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.
- [Ho *et al.*, 2022] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi,

- David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [Hu *et al.*, 2023] Jifeng Hu, Yanchao Sun, Sili Huang, SiYuan Guo, Hechang Chen, Li Shen, Lichao Sun, Yi Chang, and Dacheng Tao. Instructed diffuser with temporal condition guidance for offline reinforcement learning, 2023.
- [Imre, 2021] Baris Imre. An investigation of generative replay in deep reinforcement learning. B.S. thesis, University of Twente, 2021.
- [Janner *et al.*, 2022] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022.
- [Jiang *et al.*, 2023] Chiyu Max Jiang, Andre Cornman, Cheolho Park, Ben Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion, 2023.
- [Kandpal *et al.*, 2023] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.
- [Kang *et al.*, 2023] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning, 2023.
- [Kapelyukh *et al.*, 2022] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 8:3956–3963, 2022.
- [Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [Kiran *et al.*, 2021] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.
- [Ko *et al.*, 2023] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences, 2023.
- [Kober *et al.*, 2013] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.
- [Kong *et al.*, 2020] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.
- [Kostrikov *et al.*, 2021] Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR, 2021.
- [Kumar *et al.*, 2020] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.
- [Laskin *et al.*, 2020] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.
- [Lee and Han, 2021] Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio up-sampling. *arXiv preprint arXiv:2104.02321*, 2021.
- [Lee *et al.*, 2022] Kuang-Huei Lee, Ofir Nachum, Mengjiao Sherry Yang, Lisa Lee, Daniel Freeman, Sergio Guadarrama, Ian Fischer, Winnie Xu, Eric Jang, Henryk Michalewski, et al. Multi-game decision transformers. *Advances in Neural Information Processing Systems*, 35:27921–27936, 2022.
- [Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.
- [Li *et al.*, 2022] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.
- [Li *et al.*, 2023a] Wenhao Li, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. Hierarchical diffusion for offline decision making. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20035–20064. PMLR, 23–29 Jul 2023.
- [Li *et al.*, 2023b] Xiang Li, Varun Belagali, Jinghuan Shang, and Michael S. Ryoo. Crossway diffusion: Improving diffusion-based visuomotor policy via self-supervised learning, 2023.
- [Li *et al.*, 2023c] Zhuoran Li, Ling Pan, and Longbo Huang. Beyond conservatism: Diffusion policies in offline multi-agent reinforcement learning, 2023.
- [Li, 2023] Wenhao Li. Efficient planning with latent diffusion, 2023.
- [Liang *et al.*, 2023] Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. AdaptDiffuser: Diffusion models as adaptive self-evolving planners. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20725–20745. PMLR, 23–29 Jul 2023.
- [Lillicrap *et al.*, 2015] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

- [Liu et al., 2021] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.
- [Liu et al., 2023] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. *arXiv preprint arXiv:2302.07351*, 2023.
- [Lu et al., 2022] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022.
- [Lu et al., 2023a] Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning, 2023.
- [Lu et al., 2023b] Cong Lu, Philip J. Ball, and Jack Parker-Holder. Synthetic experience replay. In *Workshop on Reinventing Reinforcement Learning at ICLR 2023*, 2023.
- [Lugmayr et al., 2022] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.
- [Luo and Hu, 2021] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.
- [Luo et al., 2022] Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning. *arXiv preprint arXiv:2206.09328*, 2022.
- [Lyu et al., 2022] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1711–1724, 2022.
- [Mazouze et al., 2023] Bogdan Mazouze, Walter Talbott, Miguel Angel Bautista, Devon Hjelm, Alexander Toshev, and Josh Susskind. Value function estimation using conditional diffusion models for control, 2023.
- [Meng et al., 2021] Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu. Offline pre-trained multi-agent decision transformer: One big sequence model tackles all smac tasks. *arXiv preprint arXiv:2112.02845*, 2021.
- [Mishra et al., 2023] Utkarsh Aashu Mishra, Shangjie Xue, Yongxin Chen, and Danfei Xu. Generative skill chaining: Long-horizon skill planning with diffusion models. In *7th Annual Conference on Robot Learning*, 2023.
- [Mnih et al., 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [Moerland et al., 2023] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.
- [Nachum et al., 2019] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.
- [Nagabandi et al., 2018] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7559–7566. IEEE, 2018.
- [Nam et al., 2022] Taewook Nam, Shao-Hua Sun, Karl Pertsch, Sung Ju Hwang, and Joseph J Lim. Skill-based meta-reinforcement learning, 2022.
- [Neal and others, 2011] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.
- [Ng et al., 2023] Eley Ng, Ziang Liu, and Monroe Kennedy III au2. Diffusion co-policy for synergistic human-robot collaborative tasks, 2023.
- [Ni et al., 2023] Fei Ni, Jianye Hao, Yao Mu, Yifu Yuan, Yan Zheng, Bin Wang, and Zhixuan Liang. Metadiffuser: Diffusion model as conditional planner for offline meta-rl, 2023.
- [Nuti et al., 2023] Felipe Nuti, Tim Franzmeyer, and João F. Henriques. Extracting reward functions from diffusion models, 2023.
- [Oh et al., 2017] Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In *International Conference on Machine Learning*, pages 2661–2670. PMLR, 2017.
- [Pearce et al., 2023] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. Imitating human behaviour with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [Qin et al., 2020] Jiarui Qin, Weinan Zhang, Xin Wu, Jiarui Jin, Yuchen Fang, and Yong Yu. User behavior retrieval for click-through rate prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2347–2356, 2020.
- [Raab et al., 2023] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. Single motion diffusion, 2023.
- [Reed et al., 2022] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander

- Novikov, Gabriel Barth-Maron, Mai Gimenez, Yuri Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.
- [Reuss *et al.*, 2023] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies, 2023.
- [Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [Schmidhuber, 2019] Juergen Schmidhuber. Reinforcement learning upside down: Don’t predict rewards—just map them to actions. *arXiv preprint arXiv:1912.02875*, 2019.
- [Schneuing *et al.*, 2022] Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. Structure-based drug design with dvariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.
- [Schrittwieser *et al.*, 2020] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.
- [Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [Sheynin *et al.*, 2022] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.
- [Shi *et al.*, 2022] Lucy Xiaoyang Shi, Joseph J. Lim, and Youngwoon Lee. Skill-based model-based reinforcement learning, 2022.
- [Sinha *et al.*, 2021] Samarth Sinha, Ajay Mandlekar, and Animesh Garg. S4rl: Surprisingly simple self-supervision for offline reinforcement learning, 2021.
- [Song *et al.*, 2019] Yang Song, Sahaj Garg, Jiabin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation, 2019.
- [Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.
- [Song *et al.*, 2022] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.
- [Song *et al.*, 2023] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023.
- [Sootla *et al.*, 2022] Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyang Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pages 20423–20443. PMLR, 2022.
- [Sridhar *et al.*, 2023] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration, 2023.
- [Srivastava *et al.*, 2019] Rupesh Kumar Srivastava, Pranav Shyam, Filipe Mutz, Wojciech Jaskowski, and Jürgen Schmidhuber. Training agents using upside-down reinforcement learning. *arXiv preprint arXiv:1912.02877*, 2019.
- [Suh *et al.*, 2023] H. J. Terry Suh, Glen Chou, Hongkai Dai, Lujie Yang, Abhishek Gupta, and Russ Tedrake. Fighting uncertainty with gradients: Offline reinforcement learning via diffusion score matching, 2023.
- [Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [Taiga *et al.*, 2022] Adrien Ali Taiga, Rishabh Agarwal, Jesse Farebrother, Aaron Courville, and Marc G Belle-mare. Investigating multi-task pretraining and generalization in reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2022.
- [Tessler *et al.*, 2018] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.
- [Tevet *et al.*, 2022] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022.
- [Venkatraman *et al.*, 2023] Siddarth Venkatraman, Shivesh Khaitan, Ravi Tej Akella, John Dolan, Jeff Schneider, and Glen Berseth. Reasoning with latent diffusion in offline reinforcement learning, 2023.
- [Vithayathil Varghese and Mahmoud, 2020] Nelson Vithayathil Varghese and Qusay H Mahmoud. A survey of multi-task deep reinforcement learning. *Electronics*, 9(9):1363, 2020.
- [Wang *et al.*, 2023] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.
- [Watson *et al.*, 2021] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models, 2021.
- [Wen *et al.*, 2022] Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.
- [Wu *et al.*, 2019] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

- [Xian *et al.*, 2023] Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023.
- [Xiao *et al.*, 2019] Chenjun Xiao, Yifan Wu, Chen Ma, Dale Schuurmans, and Martin Müller. Learning to combat compounding-error in model-based reinforcement learning. *arXiv preprint arXiv:1912.11206*, 2019.
- [Xiao *et al.*, 2023] Wei Xiao, Tsun-Hsuan Wang, Chuang Gan, and Daniela Rus. Safediffuser: Safe planning with diffusion probabilistic models, 2023.
- [Xu *et al.*, 2022] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.
- [Xu *et al.*, 2023] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery, 2023.
- [Yang *et al.*, 2023] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators, 2023.
- [Yoneda *et al.*, 2023] Takuma Yoneda, Luzhe Sun, , Ge Yang, Bradly Stadie, and Matthew Walter. To the noise and back: Diffusion for shared autonomy, 2023.
- [Yu *et al.*, 2020] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.
- [Yu *et al.*, 2023] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspier Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.
- [Zhang *et al.*, 2022] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.
- [Zhang *et al.*, 2023a] Edwin Zhang, Yujie Lu, William Wang, and Amy Zhang. Lad: Language control diffusion: efficiently scaling through space, time, and tasks. *arXiv preprint arXiv:2210.15629*, 2023.
- [Zhang *et al.*, 2023b] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model, 2023.
- [Zhang *et al.*, 2023c] Qin Zhang, Linrui Zhang, Haoran Xu, Li Shen, Bowen Wang, Yongzhe Chang, Xueqian Wang, Bo Yuan, and Dacheng Tao. Saformer: A conditional sequence modeling approach to offline safe reinforcement learning. *arXiv preprint arXiv:2301.12203*, 2023.
- [Zheng *et al.*, 2023] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders, 2023.
- [Zhu *et al.*, 2021] Menghui Zhu, Minghuan Liu, Jian Shen, Zhicheng Zhang, Sheng Chen, Weinan Zhang, Deheng Ye, Yong Yu, Qiang Fu, and Wei Yang. Mapgo: Model-assisted policy optimization for goal-oriented tasks. *arXiv preprint arXiv:2105.06350*, 2021.
- [Zhu *et al.*, 2023] Zhengbang Zhu, Minghuan Liu, Liyuan Mao, Bingyi Kang, Minkai Xu, Yong Yu, Stefano Ermon, and Weinan Zhang. Madiff: Offline multi-agent learning with diffusion models, 2023.