DRDT3: Diffusion-Refined Decision Test-Time Training Model

Anonymous authors
Paper under double-blind review

Abstract

Decision Transformer (DT), a trajectory modelling method, has shown competitive performance compared to traditional offline reinforcement learning (RL) approaches on various classic control tasks. However, it struggles to learn optimal policies from suboptimal, reward-labelled trajectories. In this study, we explore the use of conditional generative modelling to facilitate trajectory stitching given its high-quality data generation ability. Additionally, recent advancements in Recurrent Neural Networks (RNNs) have shown their linear complexity and competitive sequence modelling performance over Transformers. We leverage the Test-Time Training (TTT) layer, an RNN that updates hidden states during testing, to model trajectories in the form of DT. We introduce a unified framework, called **D**iffusion-**R**efined **D**ecision **TTT** (DRDT3), to achieve performance beyond DT models. Specifically, we propose the Decision TTT (DT3) module, which harnesses the sequence modelling strengths of both self-attention and the TTT layer to capture recent contextual information and make coarse action predictions. DRDT3 iteratively refines the coarse action predictions through the generative diffusion model, progressively moving closer to the optimal actions. We further integrate DT3 with the diffusion model using a unified optimization objective. With experiments on multiple tasks of Gym and AntMaze in the D4RL benchmark, our DT3 model without diffusion refinement demonstrates improved performance over standard DT, while DRDT3 further achieves superior results compared to state-of-the-art conventional offline RL and DT-based methods.

1 Introduction

Reinforcement learning aims to train agents through interaction with the environment and receiving reward feedback. It has been widely applied in areas such as robotics Jia et al. (2023); Guo et al. (2023), autonomous driving Taghavifar et al. (2025); Liu et al. (2024a), energy systems Zhang et al. (2022); Bui et al. (2025), transportation systems Wang et al. (2024a); Zhai et al. (2025), and games Barros e Sá & Madeira (2025). Offline RL, also known as batch RL, eliminates the reliance on online interactions with the environment Wang et al. (2022) by directly learning policies from offline datasets composed of historical trajectories. Therefore, it is promising for applications where interactions are risky and costly. Directly applying dynamic programming-based online RL methods to offline RL problems typically learns a poor-quality policy due to overestimated out-of-distribution actions caused by distribution shift Yamagata et al. (2023). Conventional offline RL generally addresses such issues by regularizing policy to stay close to behavior policy Lyu et al. (2022) or constraining values of out-of-distribution actions Kostrikov et al. (2021); Liu et al. (2024b).

DT Chen et al. (2021), an innovative offline RL approach that formulates policy learning as a sequence modelling problem, has shown superior performance over conventional offline RL methods on some benchmarks. By leveraging the strong sequence modelling capacity of Transformer structures Vaswani et al. (2017); Radford et al. (2019), DT models trajectories autoregressively, bypassing the necessity for bootstrapping and learning a value function for a given state. Fed with the recent context, a historical subtrajectory composed of states, actions, and return-to-go (RTG) (cumulative rewards from the current step), DT predicts actions autoregressively. However, the abandonment of dynamic programming in DT makes it lack stitching ability Yamagata et al. (2023). Some recent work studies this problem. For instance, Q-learning DT Yamagata

et al. (2023) relabels RTG with a value function learned from Q-learning, enhancing DT to learn better policies from sub-optimal datasets. Elastic DT Wu et al. (2024) adjusts the context length according to the evaluated optimality of the context to achieve trajectory stitching. Our work instead addresses such a problem using generative modelling.

Recently, diffusion models Ho et al. (2020); Cao et al. (2024); Yan et al. (2024) have gained recognition for their extraordinary ability to generate high-quality complex data, such as images and texts. Additionally, recent works have also introduced the diffusion model to offline RL Janner et al. (2022); Ajay et al. (2023); Ho et al. (2020). Similar to DT, Diffuser Janner et al. (2022) also works as a trajectory generator, using expressive diffusion models instead of Transformers. Unlike existing diffusion-based offline RL methods, our work leverages diffusion models as a refinement tool.

We propose DRDT3, a framework that combines a novel DT-style trajectory modelling method and a conditional diffusion model. Specifically, we first introduce a DT3 module, which harnesses both self-attention and the TTT layer to achieve reduced complexity and improved sequence modelling ability. This module predicts coarse action representations based on recent context, which are then used as prior knowledge or conditions for a denoising diffusion probabilistic model (DDPM) Ho et al. (2020). Additionally, we present a novel gated MLP noise approximator, designed to capture essential information from both noisy input and coarse action predictions, facilitating the denoising process. To integrate the DT3 module and diffusion model effectively, we employ a unified optimization objective with two key components: an action representation term, which ensures that the DT3 module generates coarse action conditions close to the optimal action distribution, and an action refinement term that constrains the diffusion model to sample actions within the dataset distribution. This approach enables the predicted actions from DT3 to be iteratively refined with Gaussian noise through the denoising chain within the diffusion model, facilitating trajectory stitching.

The contributions of this work are summarized as follows:

(i) We introduce the DT3 model, which combines the self-attention mechanism and the TTT layer for improved action generation and enhanced performance over DT.

(i) We further propose DRDT3, a diffusion-refined DT3 algorithm that leverages coarse action representations from the DT3 module as priors and uses diffusion models to iteratively refine these actions. Additionally, we present a gated MLP noise approximator for more effective denoising. DRDT3 integrates the DT3 and diffusion model into a cohesive framework with a unified optimization objective.

(iii) Experiments on extensive tasks from the D4RL benchmark Fu et al. (2020) demonstrate the superior performance of our proposed DT3 and DRDT3 over conventional offline RL and DT-based methods.

The following sections are organized as follows: we begin with preliminaries in Section 3, covering offline RL, DT, the TTT layer, and diffusion models. The detailed methodology about our DRDT3 model are then presented in Section 4, followed by experiments conducted for evaluating our proposed models in Section 5. Section 2 provides an overview of recent related work on DT, sequence modeling and diffusion models in RL. Finally, we present conclusions and discussions in Section 6.

2 Related Work

2.1 Decision Transformer

Offline RL learns policies from historical trajectories pre-collected using other behavior policies. Different from classic offline RL Fujimoto & Gu (2021); Kostrikov et al. (2021) where dynamics programing is employed for policy optimization, DT Chen et al. (2021) formulates the policy learning as a sequence modelling problem by autoregressively generating trajectories with Generative Pre-trained Transformer (GPT)-like architectures Radford et al. (2019). Many variants of DT have been studied and shown competitive performance on some RL benchmarks. Online DT Zheng et al. (2022) equips DT with entropy regularizers, enabling online fine-tuning with efficient exploration. Prompting DT Xu et al. (2022) and Hyper-DT Xu et al. (2023) make DT easily transferable to novel tasks using a trajectory prompt and an adaptation module initialized with a hyper-network, respectively. Hierarchical DT Correia & Alexandre (2023) frees DT from specifying RTGs

while using sub-goal selection instead. Q-learning DT Yamagata et al. (2023) and Elastic DT Wu et al. (2024) improve the stitching ability of DT by introducing dynamic programing-derived value function and varying-length context, respectively. DT has also been applied to some real-world applications, such as robotics Correia & Alexandre (2023) and traffic signal control Huang et al. (2023).

2.2 Sequence Modeling

Transformer-based methods leveraging the attention mechanism have demonstrated remarkable performance across various sequence modelling domains Vaswani et al. (2017); Radford et al. (2019); Achiam et al. (2023); Peebles & Xie (2023). However, their quadratic time complexity during inference Gu & Dao (2023) poses challenges for long-sequence modelling. Recently, structured state space models (SSMs) Gu et al. (2021); Gu & Dao (2023) have introduced a compelling architecture for long-sequence processing, distinguished by their linear computational complexity. Among SSM-based methods, MAMBA has proven effective and is widely applied in various research areas Zhang et al. (2024); Wang et al. (2024b). MAMBA Gu & Dao (2023) introduces a selection mechanism to adapt SSM parameters based on input, along with parallel scanning, kernel fusion, and recomputation techniques for efficient computation. Other recent sequence modelling approaches, such as RWKV Peng et al. (2023; 2024), Gated Linear Attention (GLA) Yang et al. (2023), and xLSTM Beck et al. (2024), build on RNN architectures with enhanced expressiveness by utilizing matrix hidden states, unlike the vector hidden states of traditional RNNs. Additionally, TTT layers Sun et al. (2024) further enhance hidden state expressiveness by updating them through self-supervised learning during both training and inference. TTT-Linear and TTT-MLP, whose hidden states are a linear layer and an MLP, respectively, achieve linear time complexity with performance that is comparable to, or even surpasses, Transformers and MAMBA.

2.3 Diffusion Models in Reinforcement Learning

Diffusion models Ho et al. (2020) have been appealing for their strong capacity to generate high-dimension image or text data. In light of the expressive representation and strong multi-modal distribution modelling ability of the diffusion model, some researchers introduce the diffusion model to RL paradigms. Diffuser Janner et al. (2022) first employs diffusion models as a planner for generating trajectories in model-based offline RL, which alleviates the severe compounding errors of conventional planners. Diffusion-QL Wang et al. (2022) explores representing policy as a diffusion model and employing Q-value function guidance during training. It overcomes the over-conservatism of policies learned from conventional offline RL. Decision Diffuser Ajay et al. (2023) formulates sequential decision-making problems as conditional generative modelling and introduces constraints and skill conditions. In addition to employing diffusion models as planners or policies, SynthER Lu et al. (2024) further exploits diffusion models as a data synthesizer to augment both offline and online training data with high diversity.

Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020), a well-known diffusion model, samples data through an iterative denoising process. Our work proposes a unified framework that aims to enhance a DT-style trajectory modeling method by refining the predictions iteratively within the expressive DDPM.

3 Preliminaries

3.1 Offline Reinforcement Learning and Decision Transformer

RL trains an agent by interacting with an environment that is commonly formulated as a Markov decision process (MDP): $M = \{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma\}$. The MDP consists of state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, reward $r = \mathcal{R}(s, a)$, state transition $\mathcal{P}(s'|s, a)$, and discount factor $\gamma \in [0, 1)$ Sutton & Barto (2018). The objective of RL is to learn a policy π that maximizes the expected cumulative discounted rewards:

$$J = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}\left(s_t, a_t\right) \right], \tag{1}$$

where s_t and a_t denote state and action at time t, respectively. Offline RL typically learns a policy from offline datasets composed of historical trajectories collected by other behavior policies, eliminating the need for online interactions with the environment. A trajectory is a sequence of states, actions, and rewards up to time step T:

$$\tau = (s_1, a_1, r_1, ..., s_T, a_T, r_T). \tag{2}$$

Different from conventional dynamic programming-based offline RL methods, DT Chen et al. (2021) formulates the policy as a return-condition sequence model represented by GPT-like architectures Radford et al. (2019). Instead of generating the action based on a single given state, DT autoregressively predicts actions based on the recent context, a K-step sub-trajectory τ including historical tokens: states, actions, and RTGs $\hat{g}_t = \sum_{t'=t}^{T} r_{t'}$, where the RTG specifies the desired future return. The trajectory generation process can be formulated as

$$\tau' = (\hat{g}_{t-K+1}, s_{t-K+1}, a_{t-K+1}, \ldots, \hat{g}_t, s_t, a_t)\,. \tag{3}$$

Specifically, at time step t, DT predicts a_t based on $(\hat{g}_{t-K+1}, s_{t-K+1}, a_{t-K+1}, \ldots, \hat{g}_t, s_t)$ and repeats the procedure till the end.

3.2 Test-Time Training Layer

Due to the quadratic complexity of self-attention in processing long-sequence inputs, some recent work leverages the RNN structure for sequence modeling, given its linear complexity. However, RNNs are often constrained by their limited expressive power in hidden states. The TTT layer Sun et al. (2024) aims to address this challenge by updating the hidden state on both training and testing sequences with self-supervised learning. Concretely, each hidden state can be expressed as W_t, the weights of a model f that produces the output $z_t = f(x_t; W_t)$. The update rule for W_t is expressed as a step of gradient descent on a certain self-supervised loss ℓ:

$$W_t = W_{t-1} - \eta \nabla \ell\left(W_{t-1}; x_t\right), \tag{4}$$

where η is the learning rate and x_t is the input token. TTT layer Sun et al. (2024) sets ℓ to a reconstruction loss of x_t:

$$\ell\left(W; x_t\right) = \left\| f\left(\tilde{x}_t; W\right) - x_t \right\|^2, \tag{5}$$

where $\tilde{x}_t = \theta_K x_t$ denotes a low-rank projection corrupted input with a learnable matrix θ_K. To selectively remember information in x_t, TTT reconstructs another low-rank projection $\theta_V x_t$ instead. Therefore, the self-supervised loss becomes:

$$\ell\left(W; x_t\right) = \left\| f\left(\theta_K x_t; W\right) - \theta_V x_t \right\|^2. \tag{6}$$

To make the dimensions consistent, the output rule becomes $z_t = f(\theta_Q x_t; W_t)$. A sequence model with TTT layers consists of two training loops: the outer loop for updating the larger network and hyperparameters θ_Q, θ_K and θ_V, and the inner loop for updating W within each TTT layer. TTT further makes updating parallelizable through mini-batch TTT and accelerates computation using the dual form technique Sun et al. (2024).

3.3 Diffusion Model

Diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020) typically learns expressive representation of the data distribution $q(\mathbf{x}^0)$ from a dataset in the form $p_\theta\left(\mathbf{x}^0\right) := \int p_\theta\left(\mathbf{x}^{0:N}\right) d\mathbf{x}^{1:N}$, where latent variables $\mathbf{x}^1, ..., \mathbf{x}^N$ from timestep $i = 1$ till $i = N$ have the same dimensionality as data $\mathbf{x}^0 \sim q(\mathbf{x}^0)$. Diffusion models typically consist of two processes: diffusion or forward process, and reverse process. The forward process is defined as a Markov chain, where the data $\mathbf{x}^0 \sim q(\mathbf{x}^0)$ is gradually added with Gaussian noise following a variance schedule $\beta_1, ..., \beta_N$:

$$q\left(\mathbf{x}^i \mid \mathbf{x}^{i-1}\right) := \mathcal{N}\left(\mathbf{x}^i; \sqrt{1-\beta_i}\mathbf{x}^{i-1}, \beta_i \mathbf{I}\right). \tag{7}$$

The Gaussian transition can be alternatively formulated as $\mathbf{x}^i = \sqrt{\alpha^i}\mathbf{x}^{i-1} + \sqrt{1-\alpha^i}\boldsymbol{\epsilon}^i$, where $\alpha^i = 1 - \beta^i$ and $\boldsymbol{\epsilon}^i \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$. The simplified transition from $\mathbf{x}^0$ to $\mathbf{x}^i$ can be inferred that

$$\mathbf{x}^i = \sqrt{\bar{\alpha}^i}\mathbf{x}^0 + \sqrt{1-\bar{\alpha}^i}\boldsymbol{\epsilon}(\mathbf{x}^i, i), \tag{8}$$

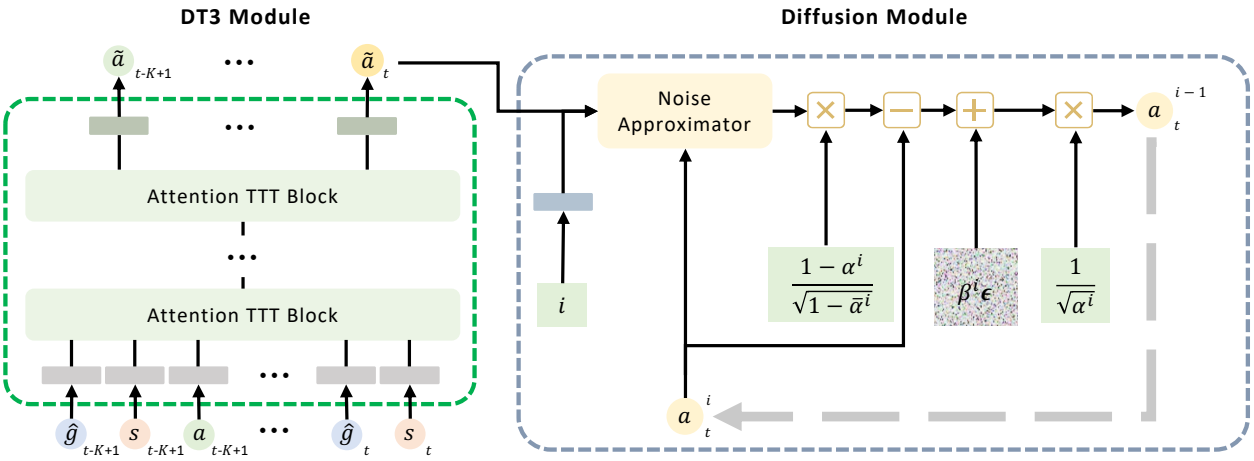

Figure 1: An illustration of the inference procedure of DRDT3. The left and right parts show the structures of DT3 and diffusion modules, respectively. During inference, the coarse action representation predicted by the DT3 module serves as the condition and is refined iteratively within the diffusion module.

where $\bar{\alpha}^i = \prod_{i=1}^{i} \alpha^i$ and $\boldsymbol{\epsilon}(\mathbf{x}^i, i) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes the cumulative noise accumulated over timesteps. The reverse process denoises a starting noise sampled from $p\left(\mathbf{x}^N\right) = \mathcal{N}\left(\mathbf{x}^N; \mathbf{0}, \mathbf{I}\right)$ back to the data distribution following a Markov chain:

$$p_\theta\left(\mathbf{x}^{i-1} \mid \mathbf{x}^i\right) := \mathcal{N}\left(\mathbf{x}^{i-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}^i, i\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}^i, i\right)\right), \tag{9}$$

where θ denotes the learnable parameters of Gaussian transitions. Generating data using diffusion models involves sampling starting noises, $\mathbf{x}^N \sim p\left(\mathbf{x}^N\right)$, and implementing reverse process from timestep $i = N$ till $i = 0$.

4 Methodology

In this part, we present the details of our DRDT3. Within the unified framework of DRDT3, a diffusion model refines the coarse action predictions from a DT3 module by injecting it as a condition and processing it using a gated MLP noise approximator. A unified objective is proposed to jointly optimize DT3 and the diffusion model.

4.1 Decision TTT

The architecture of the DT3 module is depicted in the left part of Figure 1. We modify the standard DT model by replacing Transformer blocks with our proposed Attention TTT blocks for trajectory generation. The structure of the Attention TTT block is detailed in Subsection 4.1.1. In DT3, each input token from the context is projected into the embedding space through a linear layer, and timestep embeddings are subsequently added. Input contexts with a length smaller than k are zero-padded for better generation. With the structure of DT3, actions can be generated autoregressively with known context.

4.1.1 Attention TTT Block

Given the competitive performance and linear complexity of TTT layers Sun et al. (2024) on sequence modelling, we leverage it for trajectory generation in the form of DT. Instead of using the GPT-2 model to process a sequence, we propose an Attention TTT block that exploits both self-attention and the TTT layer. As shown in the left part of Figure 2, we apply masked multi-head attention to capture correlations between input tokens, followed by an Add & Norm layer. We further employ the TTT layer to extract additional information beyond what is captured by the attention mechanism. The structure of the TTT

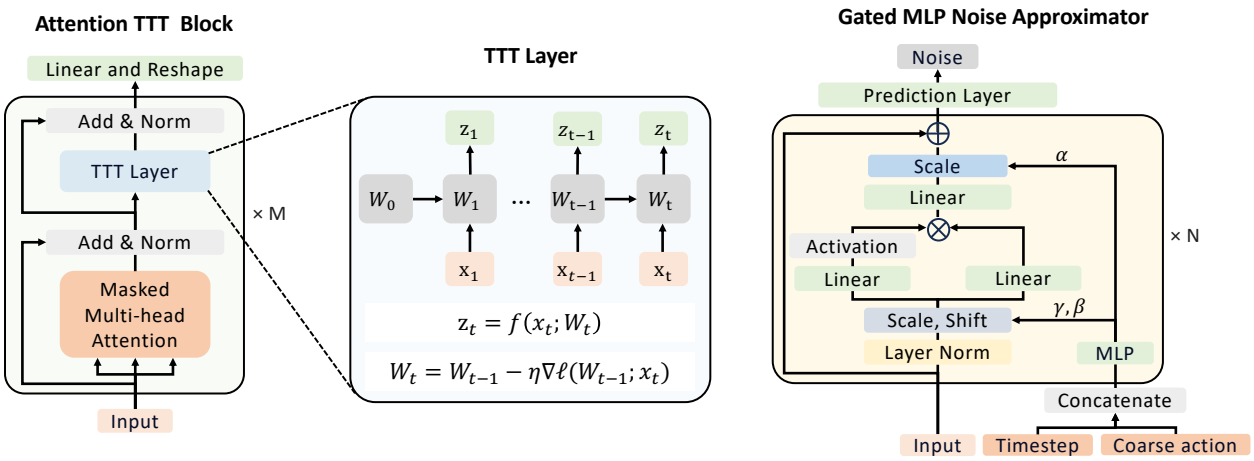

Figure 2: Structure illustration of Attention TTT block, TTT layer, and gated MLP noise approximator.

layer is illustrated in the middle part of Figure 2. In our TTT layer, we use a linear function to represent the weights W, resulting in $f(x) = Wx$, where W is a learnable square matrix. The TTT layer is also followed by an Add & Norm layer.

4.1.2 Coarse Action Prediction

Besides directly generating the final actions autoregressively using DT3, in our DRDT3, we also leverage the DT3 as one of the modules to predict a coarse representation of the action, which serves as an action condition for the subsequent diffusion model:

$$\widetilde{\boldsymbol{a}}_{-K,t} := \widetilde{\pi}_\theta(\hat{\mathbf{g}}_{-K,t}, \mathbf{s}_{-K,t}), \tag{10}$$

where θ denotes learnable parameters of DRDT3, $\widetilde{\boldsymbol{a}}_{-K,t}$ is the predicted sequence of actions, $\mathbf{s}_{-K,t}$ and $\hat{\mathbf{g}}_{-K,t}$ are the latest K-step RTGs and states, respectively. We use $\widetilde{\boldsymbol{a}}$ to denote the predicted coarse action representation for the current timestep, which is also the final term in $\widetilde{\boldsymbol{a}}_{-K,t}$.

During implementation, we initialize the RTG as the return of the entire trajectory, and the next RTG is calculated using $\hat{g}_t = \hat{g}_{t-1} - r_{t-1}$. When evaluating our DRDT3 online, we introduce an RTG scale factor to adjust the initial RTG. This is achieved by multiplying the maximum positive return from the offline trajectories by η while dividing the negative return by it. This defines the desired higher return for the upcoming evaluation.

4.2 Conditional Diffusion Policy

We adopt DDPM Ho et al. (2020), a well-known generative model, as an enhancement module to refine the coarse actions predicted from DT3. To achieve this, we extend DDPM to a conditional diffusion model and employ coarse action $\widetilde{\boldsymbol{a}}$ predicted from DT3 as a prior, as depicted in Figure 1. Consequently, the final action is sampled through a reversed process:

$$\begin{aligned} \pi_\theta(\boldsymbol{a}|\widetilde{\boldsymbol{a}}) &:= p_\theta\left(\boldsymbol{a}^{0:N}|\widetilde{\boldsymbol{a}}\right) \\ &:= p\left(\boldsymbol{a}^N\right) \prod_{i=1}^N p_\theta\left(\boldsymbol{a}^{i-1} \mid \boldsymbol{a}^i, \widetilde{\boldsymbol{a}}\right), \end{aligned} \tag{11}$$

where the final denoised sample $\boldsymbol{a}^0$ denotes the sampled action to be implemented. It is worth noting that we use subscripts $t \in \{1, ..., T\}$ to denote timesteps in trajectories while using superscripts $i \in \{1, ..., N\}$ to represent diffusion timesteps. We formulate $p_\theta\left(\boldsymbol{a}^{i-1} \mid \boldsymbol{a}^i, \widetilde{\boldsymbol{a}}\right)$ using a Gaussian transition

$$\begin{aligned} & p_\theta\left(\boldsymbol{a}^{i-1} \mid \boldsymbol{a}^i, \widetilde{\boldsymbol{a}}\right) \\ := & \mathcal{N}\left(\boldsymbol{a}^{i-1}; \boldsymbol{\mu}_\theta\left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i\right), \boldsymbol{\Sigma}_\theta\left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i\right)\right), \end{aligned} \tag{12}$$

Algorithm 1 Training of DRDT3

Initialize the parameters of policy π_θ
repeat
  Sample $\{(\hat{\mathbf{g}}_{-K}, \mathbf{s}_{-K}, \boldsymbol{a}_{-K})\}$ from $D_{offline}$
  Predict coarse action representations $\widetilde{\boldsymbol{a}}_{-K} = \widetilde{\pi}_\theta(\hat{\mathbf{g}}_{-K}, \mathbf{s}_{-K})$ and extract $\widetilde{\boldsymbol{a}}$
  Compute $\mathcal{L}_{dt3}$ by Equation 16
  Sample $i \sim \mathcal{U}(1, N)$
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Compute $\mathcal{L}_{diff}$ by Equation 15
  Update policy by minimizing $\mathcal{L}_{DRDT3} = \mathcal{L}_{diff}(\theta) + \zeta \mathcal{L}_{dt3}(\theta)$
until converged

Algorithm 2 Inference of DRDT3

Sample $\boldsymbol{a}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $i = N, ...1$ **do**
  Predict coarse action representation $\widetilde{\boldsymbol{a}}$ with the DT3 module
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $i > 1$ else $\boldsymbol{\epsilon} = 0$
  $\boldsymbol{a}^{i-1} = \frac{1}{\sqrt{\alpha^i}} \left(\boldsymbol{a}^i - \frac{1-\alpha^i}{\sqrt{1-\bar{\alpha}^i}} \boldsymbol{\epsilon}_\theta \left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i \right) \right) + \beta^i \boldsymbol{\epsilon}$
end for
Return $\boldsymbol{a}^0$ as the selected action

where the covariance matrix is set as $\boldsymbol{\Sigma}_\theta \left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i \right) = \beta^i \mathbf{I}$. Following DDPM Ho et al. (2020), we can infer the mean as

$$\boldsymbol{\mu}_\theta \left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i \right) := \frac{1}{\sqrt{\alpha^i}} \left(\boldsymbol{a}^i - \frac{1-\alpha^i}{\sqrt{1-\bar{\alpha}^i}} \boldsymbol{\epsilon}_\theta \left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i \right) \right), \tag{13}$$

where the noise prediction model $\boldsymbol{\epsilon}_\theta$ is represented using neural networks. Therefore, during inference, as shown in Algorithm 2, we first sample $\boldsymbol{a}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The coarse action can be refined iteratively following the formulation below till we get the final action $\boldsymbol{a}^0$

$$\boldsymbol{a}^{i-1} = \frac{1}{\sqrt{\alpha^i}} \left(\boldsymbol{a}^i - \frac{1-\alpha^i}{\sqrt{1-\bar{\alpha}^i}} \boldsymbol{\epsilon}_\theta \left(\boldsymbol{a}^i, \widetilde{\boldsymbol{a}}, i \right) \right) + \beta^i \boldsymbol{\epsilon}, \tag{14}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ when $i > 1$ while $\boldsymbol{\epsilon} = 0$ when $i = 1$ for better sample quality Ho et al. (2020). We choose to predict the noise $\boldsymbol{\epsilon}$ using neural networks. Therefore, according to DDPM Ho et al. (2020), the noise approximator $\boldsymbol{\epsilon}_\theta$ can be optimized with a simplified loss function

$$\mathcal{L}_{diff} := \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{a}, \widetilde{\boldsymbol{a}}, i} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\sqrt{\bar{\alpha}^i} \boldsymbol{a} + \sqrt{1-\bar{\alpha}^i} \boldsymbol{\epsilon}, \widetilde{\boldsymbol{a}}, i \right) \right\|^2 \right], \tag{15}$$

where $i \sim \mathcal{U}(1, N)$ and $\boldsymbol{a} \sim \mathcal{D}_{offline}$.

4.3 Gated MLP Noise Approximator

To fully integrate coarse action predictions into the diffusion model and improve noise prediction accuracy, we introduce a gated MLP noise approximator, as depicted in the right part of Figure 2. The gated MLP noise approximator utilizes adaptive layer normalization (adaLN) Peebles & Xie (2023) to capture condition-specific information. It specifically learns the scale and shift parameters, γ and β, for layer normalization, as well as a scaling parameter α based on the concatenated condition of the diffusion timestep and coarse action prediction. Additionally, we employ a gated MLP block Dauphin et al. (2017); De et al. (2024) to extract multi-granularity information from the input, enhancing the model's noise reconstruction capabilities. This architecture splits the output from the Scale & Shift layer into two separate pathways, each expanding

the feature dimensionality by a factor of M using a linear layer. The GeLU Hendrycks & Gimpel (2016) non-linearity is applied in one branch, after which the two streams are combined through element-wise multiplication. A final linear layer reduces back the dimensionality, and a residual connection is incorporated to facilitate better gradient flow.

4.4 Unified optimization objective

The diffusion loss $\mathcal{L}_{diff}(\theta)$ facilitates refinement for the coarse action representation $\widetilde{\boldsymbol{a}}$ and behavior cloning towards the dataset, while it fails to constrain the DT3 model, potentially leading to deviations in coarse action predictions from the data distribution. To alleviate such issues and jointly optimize DT3 and the diffusion model, we propose a unified optimization objective. This involves introducing a DT3 loss as additional guidance during the training of the diffusion model

$$\mathcal{L}_{dt3} := \frac{1}{K\boldsymbol{a}_{\max}} \mathbb{E}_{(\boldsymbol{a},\mathbf{s},\hat{\mathbf{g}})\sim\mathcal{D}_{offline}} \left[\|\boldsymbol{a}_{-K} - \widetilde{\pi}_\theta(\hat{\mathbf{g}}_{-K}, \mathbf{s}_{-K})\|_1 \right], \tag{16}$$

where $\boldsymbol{a}_{-K}$ is the ground truth of the latest K-step actions, and $\boldsymbol{a}_{\max}$, the maximum action in the action space, serves as a scaling factor to render the L_1 loss dimensionless. It is worth noting that the diffusion model only conditions on the last term $\widetilde{\boldsymbol{a}}$ within the sequence $\widetilde{\boldsymbol{a}}_{-K,t}$, whereas the DT3 loss is computed for the entire sequence. The final unified loss for DRDT3 is a linear combination of the DT3 loss (action representation term) and diffusion loss (action refinement term)

$$\mathcal{L}_{DRDT3} := \mathcal{L}_{diff}(\theta) + \zeta \mathcal{L}_{dt3}(\theta). \tag{17}$$

where ζ is a loss coefficient for balancing the two loss terms. Algorithm 1 and 2 illustrate the training and inference procedures of DRDT3.

4.5 Implementation Details

When implementing our proposed DRDT3, we train it for 50 epochs with 2000 gradient updates per epoch. The learning rate and batch size are designated as 0.0003 and 2048, respectively. To proceed with historical subtrajectories with DT3, we set the context length as 6. The Attention TTT block used in the DT3 module consists of 1-layer self-attention and 1-layer TTT with embedding dimensions of 128. We use a linear layer to output a deterministic coarse representation of the action. To enhance computational efficiency without sacrificing performance, the number of diffusion timesteps is set to 5. We set the variance schedule according to Variance Preserving (VP) SDE Xiao et al. (2021):

$$\beta^i = 1 - \alpha^i = 1 - e^{-\beta_{\min}\left(\frac{1}{N}\right) - 0.5(\beta_{\max}-\beta_{\min})\frac{2i-1}{N^2}}. \tag{18}$$

Most of the hyperparameters are selected through the Optuna hyperparameter optimization framework Akiba et al. (2019).

5 Experiments

We conduct experiments to evaluate our proposed DRDT3 on the commonly used D4RL benchmark Fu et al. (2020) using an AMD Ryzen 7 7700X 8-Core Processor with a single NVIDIA GeForce RTX 4080 GPU.

5.1 Experimental Settings

5.1.1 Tasks and Datasets

We adopt four Mujoco locomotion tasks from Gym, which are HalfCheetah, Hopper, Walker2D, and Ant, and a goal-reaching task, AntMaze. For the Gym tasks, dense rewards are provided, and we use three different dataset settings from D4RL: Medium (m), Medium-Replay (mr), and Medium-Expert (me). Specifically, Medium datasets contain one million samples collected using a behavior policy which achieves 1/3 scores

Table 1: Normalized scores of different offline RL methods on Gym and AntMaze tasks from D4RL benchmark. Results of DRDT3 represent the mean and variance across three seeds. The best scores are highlighted in bold.

Dataset	BC	TD3BC	CQL	IQL	Diffuser	DT	ConDT	QDT	EDT	DRDT3 (ours)
halfcheetah-m	42.6	**48.3**	44.0	47.4	42.8	42.6	43.0	42.3	42.5	44.1±1.3
hopper-m	52.9	59.3	58.5	66.3	74.3	67.6	74.5	66.5	63.5	**92.7±8.8**
walker2d-m	75.3	83.7	72.5	78.3	79.6	74	72.8	67.1	72.8	**84.2±1.8**
halfcheetah-mr	36.6	44.6	**45.5**	44.2	37.7	36.6	40.9	35.6	37.8	42.3±0.5
hopper-mr	18.1	60.9	95.0	94.7	93.6	82.7	**95.1**	52.1	89.0	92.7±2.1
walker2d-mr	26.0	81.8	77.2	73.9	70.6	66.6	72.5	58.2	74.8	**85.0±1.8**
halfcheetah-me	55.2	90.7	91.6	86.7	88.9	86.8	92.6	79.0	89.1	**94.2±2.1**
hopper-me	52.5	98.0	105.4	91.5	103.3	107.6	110.3	94.2	108.7	**112.5±1.5**
walker2d-me	107.5	**110.1**	108.8	109.6	106.9	108.1	109.1	101.7	106.2	104.5±6.4
Average	51.9	75.3	77.6	77	77.5	74.7	79.0	66.3	76.0	**83.6**
ant-m	-	-	-	99.9	-	93.6	-	-	97.9	**103.3±5.8**
ant-mr	-	-	-	91.2	-	89.1	-	-	92.0	**96.7±5.1**
Average	-	-	-	95.6	-	91.4	-	-	95.0	**100.0**
antmaze-umaze	54.6	78.6	74.0	**87.5**	-	59.2	-	67.2	-	76.6±1.9
antmaze-umaze-diverse	45.6	71.4	**84.0**	62.2	-	53	-	62.1	-	82.5±6.3
Average	50.1	75.0	79.0	74.9	-	56.1	-	64.7	-	**79.6**

of an expert policy; Medium-Reply datasets contain the replay buffer during training a policy till reaching the Medium score; Medium-Expert datasets consist of two million samples evenly generated from a medium agent and an expert agent. The AntMaze task aims to move an ant robot to reach a target location with sparse rewards (1 for reaching the goal, otherwise 0). Following experimental settings in a DT variant Zheng et al. (2022), we adopt the umaze and umaze-diverse datasets for evaluation.

5.1.2 Baselines

To verify the effectiveness of our proposed DRDT3, we compare it with several baseline methods.

- **Behavior Cloning (BC)**: A supervised learning approach that trains a policy to mimic the behavior of expert demonstrations by learning from state-action pairs.
- **TD3+BC** Fujimoto & Gu (2021): An offline RL algorithm that adapts the TD3 method Fujimoto et al. (2018) by incorporating a behavior cloning regularization term into the policy objective. This modification aims to prevent the policy from generating actions that deviate significantly from the offline dataset distribution.
- **Conservative Q-Learning (CQL)** Kumar et al. (2020): An offline RL algorithm that penalizes the Q-values associated with out-of-distribution actions, thereby encouraging more conservative Q-function estimates.
- **Implicit Q-Learning (IQL)** Kostrikov et al. (2021): IQL approximates the upper expectile of Q-value distributions, allowing it to learn policies focusing on actions within the dataset distribution.
- **Diffuser** Janner et al. (2022): A diffusion-based algorithm that adopts diffusion models as a planner for generating trajectories in the form of model-based offline RL.
- **DT** Chen et al. (2021): DT formulates the policy learning as a sequence modelling problem by autoregressively generating trajectories using GPT models.
- **Contrastive DT (ConDT)** Konan et al. (2023): A DT variant that harnesses the power of contrastive representation learning and return-dependent transformations to cluster the input embeddings based on their associated returns.
- **Q-learning DT (QDT)** Yamagata et al. (2023): A DT variant that improves the stitching ability of DT by introducing dynamic programming-derived value function.
- **Elastic DT (EDT)** Wu et al. (2024): A DT variant that dynamically adjusts the length of the input context according to the quality of the previous trajectory.

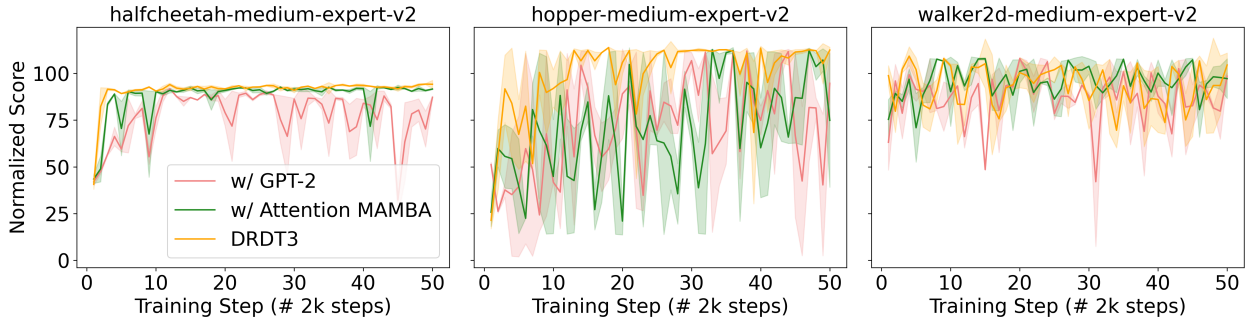

Figure 3: Ablation study on sequence model.

The best scores reported in papers of these baselines are adopted. For ConDT, we calculate the normalized score according to the returns reported in Konan et al. (2023).

Table 2: Performance comparison between DT, DT3 and DRDT3 on Gym locomotion tasks. The underlined values indicate performance better than DT, and bold values are the best results. 'hc': halfcheetah; 'hp': hopper; 'wk': walker2d.

	hc-m	hp-m	wk-m	hc-mr	hp-mr	wk-mr	hc-me	hp-me	wk-me	Average
DT	**42.6**	67.6	74.0	36.6	**82.7**	**66.6**	86.8	107.6	**108.1**	74.7
DT3 (ours)	42.6±0.7	73.0±6.6	79.3±1.1	39.9±0.4	75.9±4.4	63.9±2.8	93.4±2.7	109.9±1.8	104.3±8.0	75.8
DRDT3 (ours)	**44.1±1.3**	**92.7±8.8**	**84.2±1.8**	**42.3±0.5**	**92.7±2.1**	**85.0±1.8**	**94.2±2.1**	**112.5±1.5**	104.5±6.4	**83.6**

5.2 Performance Comparison

Table 1 presents the normalized scores Fu et al. (2020) of our DRDT3 and baselines on several tasks from the D4RL benchmark. As we can see, DRDT3 significantly improves the average scores on both Gym and AntMaze tasks. Specifically, DRDT3 outperforms other offline RL methods, including the state-of-the-art DT variants, on 7 out of 11 Gym tasks. For the other 4 Gym tasks, DRDT3 is still comparable over the best baselines. Particularly, the Medium-Expert dataset is not reported for the Ant task, given the lack of comparing baselines evaluated on it. For AntMaze tasks, DRDT3 shows comparable performance over the best baselines for each single task while achieving the best overall results. When compared with the vanilla DT, our DRDT3 presents significant improvements on 12 out of 13 tasks, especially on tasks using Medium and Medium-Replay datasets with sub-optimal trajectories, which demonstrates the enhanced stitching ability of our DRDT3 over DT.

5.3 Ablation Study

We implement ablation studies for some components to demonstrate their effectiveness.

5.3.1 Ablation on Sequence Model

We compare our DRDT3 framework with two variants: one where the sequence model is replaced by a GPT-2 model Radford et al. (2019) and another where the TTT layer is replaced by an MAMBA layer Gu & Dao (2023). The GPT-2 model relies entirely on the attention mechanism, while MAMBA is a selective state space model based on the RNN structure, where parameters are dynamically dependent on the inputs.

Table 3: Performance comparison between DT3 and DT on Ant and AntMaze tasks.

	ant-m	ant-mr	Average	antmaze-umaze	antmaze-umaze-diverse	Average
DT	93.6	**89.1**	91.4	59.2	**53.0**	56.1
DT3 (ours)	96.7±7.3	89.0±5.2	92.9	62.4±3.3	53.0±4.8	57.7
DRDT3 (ours)	**103.3±5.8**	**96.7±5.1**	**100.0**	**76.6±1.9**	**82.5±6.3**	**79.6**

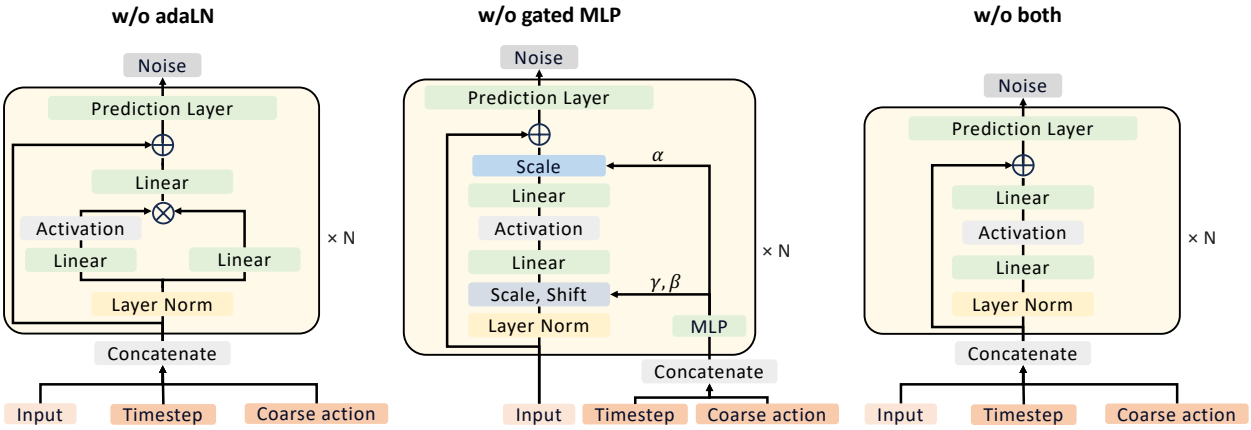

Figure 4: Structure illustration of three noise approximators. **Left**: w/o adaLN; **Middle**: w/o gated MLP; **Right**: w/o both.

For the two variants w/GPT-2 and w/ Attention MAMBA, we use one block per model as our DRDT3. To ensure a fair comparison, we adjust the number of layers in both the GPT-2 and MAMBA variants to match the model size of our DRDT3. Specifically, the total number of learnable parameters for our DRDT3, the variant w/GPT-2, and the variant w/ Attention MAMBA are 235K, 218K, and 329K, respectively. We present the structures of three noise approximations adopted for comparison in Figure 4.

As shown in Figure 3, our DRDT3 with Attention TTT blocks outperforms the other variants across three Medium-Expert tasks. The variant with Attention MAMBA shows slight performance degradation and lacks the stability of DRDT3. The variant with GPT-2 performs the worst across all tasks, which might be attributed to the typical reliance on larger model sizes of GPT-2. When constrained to a similar size as the other two variants, GPT-2 experiences performance degradation. These results indicate that our proposed DT3 module with Attention TTT blocks enhances performance without increasing model size. It is worth noting that our DRDT3 does not show improved computation efficiency over other variants, but the final performance is promising. This may be due to the use of short-range context as the sequence input, whereas the MAMBA and TTT layers are optimized for more efficient long-sequence processing.

5.3.2 Ablation on Diffusion

To test whether the diffusion module plays a significant role in action refinement and verify the effectiveness of our DT3 module, we train a DT3 model along with the same settings as DRDT3 by minimizing the $\mathcal{L}_{dt3}$ in Equation 16. Table 2 and 3 show the performance comparison between DT and our DT3 model on Gym and AntMaze tasks. As we can see, our proposed DT3 trained independently can outperform DT on 6 out of 9 Gym locomotion tasks and all AntMaze tasks. This demonstrates the superiority of our proposed DT3 with the Attention TTT blocks over the vanilla DT with only the attention mechanism. Furthermore, our proposed DRDT3, which enhances DT3 through a diffusion module, further boosts performance across all tasks with different datasets, achieving average improvements of 10.3%, 7.6%, and 38.0% on Gym locomotion, Ant, and AntMaze tasks, respectively. This demonstrates the substantial impact of incorporating diffusion refinement in DRDT3.

5.3.3 Ablation on Noise Approximator

We also verify the effectiveness of our gated MLP noise approximator by comparing DRDT3 with several variants: one where the gated MLP structure is replaced by a simple MLP, another where the adaLN is replaced with in-context conditioning Peebles & Xie (2023), and a third variant that removes both structures. As shown in Table 4, the removal of adaLN causes a slight performance degradation, while removing the gated MLP results in a significant reduction in performance.

Table 4: Ablation studies on the *noise approximator* (left side) and *unified loss function* (right side).

Dataset	w/o adaLN	w/o gated mlp	w/o both	DRDT3 (ours)	w/o $\mathcal{L}_{dt3}$	w/ L2 $\mathcal{L}_{dt3}$
halfcheetah-m	43.4±1.6	44.0±1.5	43.8±1.2	**44.1±1.3**	**44.1±1.4**	42.5±1.3
hopper-m	91.3±4.9	77.3±5.2	75.4±6.0	**92.7±8.8**	81.2±8.2	92.0±5.2
walker2d-m	84.0±1.0	83.4±1.1	82.5±1.8	**84.2±1.8**	80.5±2.1	79.8±1.2
halfcheetah-mr	42.7±0.8	42.8±0.9	**44.4±0.8**	42.3±0.5	42.7±0.9	38.2±0.9
hopper-mr	**92.7±1.3**	86.9±1.5	85.9±5.1	**92.7±2.1**	84.2±5.4	91.5±1.7
walker2d-mr	85.5±2.8	82.2±2.7	80.7±0.9	85.0±1.8	85.1±1.3	**86.0±2.9**
halfcheetah-me	93.8±2.5	89.8±3.1	90.3±2.5	**94.2±2.1**	92.3±2.6	93.3±3.0
hopper-me	111.9±1.7	106.6±1.5	105.9±1.3	**112.5±1.5**	105.1±1.6	110.3±1.5
walker2d-me	102.6±1.3	101.1±1.1	98.4±5.2	**104.5±6.4**	100.7±5.6	103.0±1.3
Average	83.1	79.3	78.6	**83.6**	79.5	81.8

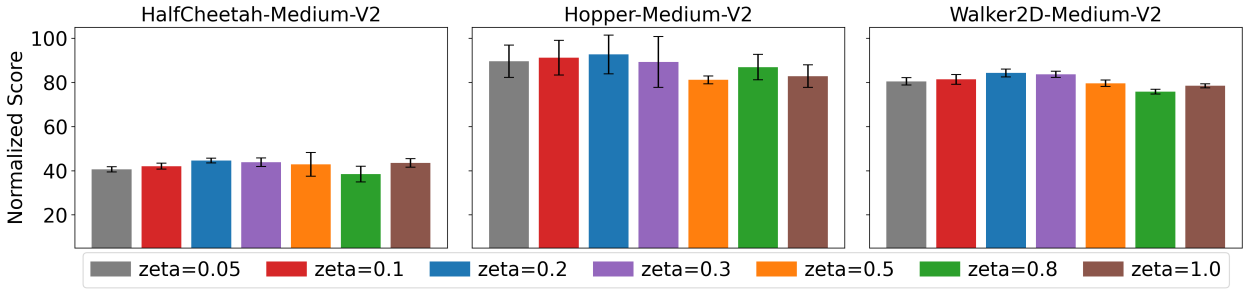

Figure 5: Sensitivity analysis on the loss coefficient ζ. We empirically choose $\zeta = 0.2$ according to the performance comparison.

5.3.4 Ablation on Loss Function

To verify the effectiveness of our proposed unified optimization objective, we compare our loss function with the vanilla diffusion loss function (w/o $\mathcal{L}_{dt3}$) and a unified loss function incorporating an L2-based DT3 loss term (w/ L2 $\mathcal{L}_{dt3}$). As shown in Table 4, the variant w/o $\mathcal{L}_{dt3}$ shows worse performance than the variant with an extra L2-based DT3 constraint. Our DRDT3, utilizing an L1-based DT3 loss term instead, can further enhance performance beyond the L2 variant. These results demonstrate that incorporating DT3 loss into the final loss improves performance, with the L1 $\mathcal{L}_{dt3}$ offering additional gains over the L2 variant.

5.3.5 Sensitivity Analysis on Loss Coefficient

We further vary the loss coefficient ζ across different values to select the best one empirically. Figure 5 illustrates the normalized scores of our DRDT3 with varying ζ and trained on three Gym tasks with Medium datasets. We empirically set $\zeta = 0.2$ based on the results.

6 Conclusion and Discussion

This work introduces the DRDT3 algorithm to enhance the stitching ability of the DT-based method. We propose a novel DT3 module with attention TTT blocks to process historical information and predict a better coarse action representation. The diffusion module then conditions this coarse representation and iteratively refines the action using a novel gated MLP noise approximator, improving performance on sub-optimal datasets. To achieve joint learning of the DT3 and diffusion modules, we introduce a unified optimization objective, consisting of both an action representation term and an action refinement term. Experiments on a variety of tasks with different dataset configurations from the D4RL benchmark demonstrate the effectiveness and superior performance of DRDT3 compared to conventional offline RL methods and DT-based approaches. Experiment results also demonstrate that DT3 can outperform the vanilla DT. In future work, we plan to explore the potential of extending DRDT3 for online adaptation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2023.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Gabriel Caldas Barros e Sá and Charles Andrye Galvão Madeira. Deep reinforcement learning in real-time strategy games: a systematic literature review. *Applied Intelligence*, 55(3):243, 2025.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

Van-Hai Bui, Sina Mohammadi, Srijita Das, Akhtar Hussain, Guilherme Vieira Hollweg, and Wencong Su. A critical review of safe reinforcement learning strategies in power and energy systems. *Engineering Applications of Artificial Intelligence*, 143:110091, 2025.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

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.

André Correia and Luís A Alexandre. Hierarchical decision transformer. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1661–1666. IEEE, 2023.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models, feb. 2024. *URL http://arxiv.org/abs/2402.19427 v1*, 2024.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Siyuan Guo, Lixin Zou, Hechang Chen, Bohao Qu, Haotian Chi, S Yu Philip, and Yi Chang. Sample efficient offline-to-online reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xingshuai Huang, Di Wu, and Benoit Boulet. Traffic signal control using lightweight transformers: An offline-to-online rl approach. *arXiv preprint arXiv:2312.07795*, 2023.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Yuanyuan Jia, Pedro Miguel Uriguen Eljuri, and Tadahiro Taniguchi. A bayesian reinforcement learning method for periodic robotic control under significant uncertainty. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 384–391. IEEE, 2023.

Sachin G Konan, Esmaeil Seraj, and Matthew Gombolay. Contrastive decision transformers. In *Conference on Robot Learning*, pp. 2159–2169. PMLR, 2023.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pp. 5774–5783. PMLR, 2021.

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.

Jia Liu, Yunduan Cui, Jianghua Duan, Zhengmin Jiang, Zhongming Pan, Kun Xu, and Huiyun Li. Reinforcement learning-based high-speed path following control for autonomous vehicles. *IEEE Transactions on Vehicular Technology*, 2024a.

Jie Liu, Yinmin Zhang, Chuming Li, Yaodong Yang, Yu Liu, and Wanli Ouyang. Adaptive pessimism via target q-value for offline reinforcement learning. *Neural Networks*, 180:106588, 2024b.

Cong Lu, Philip Ball, Yee Whye Teh, and Jack Parker-Holder. Synthetic experience replay. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Hamid Taghavifar, Chuan Hu, Chongfeng Wei, Ardashir Mohammadzadeh, and Chunwei Zhang. Behaviorally-aware multi-agent rl with dynamic optimization for autonomous driving. *IEEE Transactions on Automation Science and Engineering*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Maonan Wang, Xi Xiong, Yuheng Kan, Chengcheng Xu, and Man-On Pun. Unitsa: A universal reinforcement learning framework for v2x traffic signal control. *IEEE Transactions on Vehicular Technology*, 2024a.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. Is mamba effective for time series forecasting? *arXiv preprint arXiv:2403.11144*, 2024b.

Yueh-Hua Wu, Xiaolong Wang, and Masashi Hamaya. Elastic decision transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pp. 24631–24645. PMLR, 2022.

Mengdi Xu, Yuchen Lu, Yikang Shen, Shun Zhang, Ding Zhao, and Chuang Gan. Hyper-decision transformer for efficient online policy adaptation. *arXiv preprint arXiv:2304.08487*, 2023.

Taku Yamagata, Ahmed Khalil, and Raul Santos-Rodriguez. Q-learning decision transformer: Leveraging dynamic programming for conditional sequence modelling in offline rl. In *International Conference on Machine Learning*, pp. 38989–39007. PMLR, 2023.

Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8239–8249, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Ziyang Zhai, Ruru Hao, Boyang Cui, and Siyi Wang. Hgat and multi-agent rl-based method for multi-intersection traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, Ziyang Wang, and Zi Ye. A survey on visual mamba. *Applied Sciences*, 14(13):5683, 2024.

Huiliang Zhang, Di Wu, and Benoit Boulet. Metaems: A meta reinforcement learning-based control framework for building energy management system. *arXiv preprint arXiv:2210.12590*, 2022.

Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In *international conference on machine learning*, pp. 27042–27059. PMLR, 2022.