Prompt Tuning with Diffusion for Few-Shot Pre-trained Policy Generalization

Extended Abstract

Shengchao Hu
Shanghai Jiao Tong University
Shanghai, China
charles-hu@sjtu.edu.cn

Wanru Zhao
University of Cambridge
Cambridge, United Kingdom
wz341@cam.ac.uk

Weixiong Lin
Shanghai Jiao Tong University
Shanghai, China
wx_lin@sjtu.edu.cn

Li Shen
Sun Yat-sen University
Shenzhen, China
mathshenli@gmail.com

Ya Zhang
Shanghai Jiao Tong University
Shanghai, China
ya_zhang@sjtu.edu.cn

Dacheng Tao
Nanyang Technological University
Singapore, Singapore
dacheng.tao@ntu.edu.sg

ABSTRACT

Offline reinforcement learning (RL) methods harness previous experiences to derive an optimal policy, forming the foundation for pre-trained large-scale models (PLMs). When adapting to novel tasks, PLMs leverage expert trajectories as prompts to accelerate adaptation. While various prompt-tuning techniques aim to improve prompt quality, their effectiveness is often limited by initialization constraints, restricting exploration and potentially leading to sub-optimal solutions. To eliminate dependence on the initial prompt, we reframe prompt-tuning as conditional generative modeling, where prompts are generated from random noise. Our proposed Prompt Diffuser employs a conditional diffusion model to generate high-quality prompts. Central to our framework is trajectory reconstruction and the seamless integration of downstream task guidance during training. Experimental results validate Prompt Diffuser's effectiveness, demonstrating strong performance in meta-RL tasks.

KEYWORDS

Diffusion Model; Offline Reinforcement Learning; Prompt Tuning; Meta Learning; Task Adaptation

ACM Reference Format:
Shengchao Hu, Wanru Zhao, Weixiong Lin, Li Shen, Ya Zhang, and Dacheng Tao. 2025. Prompt Tuning with Diffusion for Few-Shot Pre-trained Policy Generalization: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

In RL, offline decision-making is crucial for deriving optimal policies from behavior-policy-generated trajectories without real-time environmental interactions. Recent advances [4, 5, 8, 11] leverage transformer-based architectures and sequence modeling to tackle multi-task offline RL. By employing prompt-tuning, these methods achieve efficient adaptation to target tasks while fine-tuning only a small subset of parameters. Despite its efficiency, prompt-tuning is highly sensitive to initialization [6, 9]. When initialized with a random prompt, the PLMs may explore only a constrained region, causing optimization to converge to a suboptimal prompt [7].

To mitigate reliance on the quality of initial prompts, we reframe prompt-tuning as a conditional generative modeling problem, where prompts are generated from random noise. This approach eliminates the need for expert-curated prompts, with the quality of generated prompts governed by the generative model's parameters, which may incorporate prior knowledge via pre-training on a fixed dataset. However, in few-shot meta-learning settings, the limited availability of offline target-task data necessitates rapid model adaptability. Additionally, these offline datasets often lack expert-level quality, requiring the generative model to produce prompts that surpass the fine-tuning data rather than merely replicating its distribution. Given the semantic sensitivity of trajectory prompts, even minor perturbations can induce significant shifts in meaning [6], underscoring the need for precision in prompt generation.

To tackle these challenges, we propose Prompt Diffuser (see Figure 1), a novel algorithm that employs a conditional diffusion model to generate high-quality prompts. Our framework establishes a trajectory representation and conditions the generative model on returns, ensuring precision and facilitating rapid adaptation to new tasks. However, optimizing Prompt Diffuser solely with the DDPM loss yields performance comparable to the original dataset [10, 13]. To enhance prompt quality, we integrate downstream task guidance into the reverse diffusion process. By applying gradient projection techniques, we incorporate this guidance without compromising the diffusion model's overall performance, achieved by projecting the guidance loss gradient onto the orthogonal complement of the diffusion loss subspace. This paper provides a preliminary view of the problem, please see [7] for a more extensive study.

2 METHODS

We formulate prompt-tuning as a standard conditional generative modeling (GM) problem:

$$\max_{\theta} \mathbb{E}_{s_0 \sim d_0}\left[\sum_{t=1}^{T} \mathcal{R}(s_t, \mathrm{PLM}(\tau^*_{\text{prompt}}, s_{0:t}, a_{0:t-1},))\right], \tag{1}$$

$$\text{s.t. } \tau^*_{\text{prompt}} \sim \mathrm{GM}_\theta(\tau^*_{\text{initial}} \mid C), \tag{2}$$

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

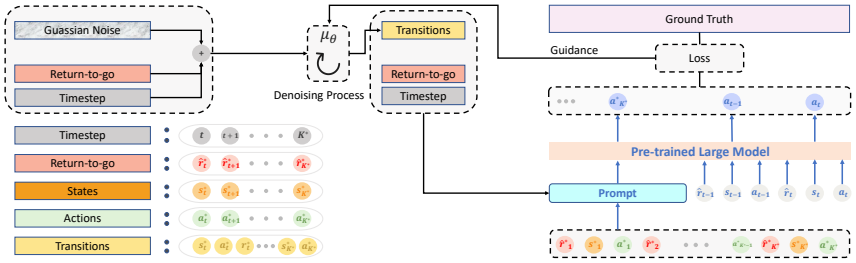

Figure 1: Overall architecture of Prompt Diffuser. Diffuser samples transitions conditioned on the return-to-go and timestep tokens, which construct a prompt for the PLM. The loss between predicted and actual actions guides the denoising process, enhancing the quality of the generated prompts.

where T denotes the maximum number of interactions, and C represents conditioning factors such as return, constraints, or demonstrated skills. We adopt Prompt-DT [15] as the PLM and an MLP-based diffusion model [2, 12] as GM.

The trajectory is constructed via a conditional diffusion process:

$$q(x^{n+1}(\tau^*) \mid x^n(\tau^*)), \quad p_\theta(x^{n-1}(\tau^*)|x^n(\tau^*), y(\tau^*)), \tag{3}$$

where q denotes the forward noising process, while p_θ represents the reverse denoising process. The initial $x^0(\tau^*)$ encodes states, actions, and rewards:

$$x^0(\tau^*) := \begin{bmatrix} s_t^* & s_{t+1}^* & \cdots & s_{t+K^*-1}^* \\ a_t^* & a_{t+1}^* & \cdots & a_{t+K^*-1}^* \\ r_t^* & r_{t+1}^* & \cdots & r_{t+K^*-1}^* \end{bmatrix}, \tag{4}$$

with the condition:

$$y(\tau^*) := \begin{bmatrix} \hat{r}_t^* & \hat{r}_{t+1}^* & \cdots & \hat{r}_{t+K^*-1}^* \\ t & t+1 & \cdots & t+K^*-1 \end{bmatrix}, \tag{5}$$

where $y(\tau^*)$ contains the returns-to-go $\hat{r}_t^* = \sum_{t'=t}^{T} r_{t'}^*$ and timesteps.

The training objective consists of two parts. First, we adopt denoising diffusion probabilistic modeling (DDPM) [2], incorporating additional conditions into the reverse diffusion process p_θ, parameterized by the noise model ϵ_θ:

$$\begin{aligned} L_{DM} =& \mathbb{E}_{n\sim\mathcal{U},\tau^*\sim\mathcal{D}^*,\epsilon_n\sim\mathcal{N}(0,I)}[||\epsilon_n - \epsilon_\theta(\sqrt{\bar{\alpha}_n}x^0(\tau^*) \\ &+ \sqrt{1-\bar{\alpha}_n}\epsilon_n, y(\tau^*), n)||^2], \end{aligned} \tag{6}$$

where $\mathcal{U}$ is a uniform distribution over the discrete set as $\{1, \ldots, N\}$ and $\mathcal{D}^*$ is the dataset collected by behavior policy π_β.

To enhance prompt quality, we integrate downstream task guidance into the reverse diffusion chain, optimizing prompts for improved performance in downstream tasks. The downstream task loss [1, 3] is defined as:

$$L_{DT} = \mathbb{E}_{\tau_i^{input}\sim\mathcal{D}_i}\left[\frac{1}{K}\sum_{k=1}^{K}(a_{i,k}^+ - \text{PLM}(x^0(\tau_i^*), y(\tau_i^*), \tau_{i,k}^+))^2\right], \tag{7}$$

where τ^+ represents the most recent historical trajectory.

To balance diffusion loss L_{DM} and downstream task loss L_{DT}, we analyze their correlation using gradient projection. Let $S_{DM}^\perp = \text{span}\{B\} = \text{span}\{[u_1, \ldots, u_M]\}$ represent the subspace spanned by $\nabla L_{DM}^\perp$, where B constitutes the bases for $S_{DM}^\perp$ and $(\cdot)^\perp$ denotes the orthogonal space (consisting of a total of M bases extracted from $\nabla L_{DM}^\perp$). The projection of any matrix A onto $S_{DM}^\perp$ is given by:

$$\text{Proj}_{S_{DM}^\perp}(A) = ABB^\top, \tag{8}$$

where $(\cdot)^\top$ is the matrix transpose. Using this projection,, the final update gradient can be:

$$\nabla L = \begin{cases} \mathbf{g}_{DM} + \lambda \cdot \text{Proj}_{S_{DM}^\perp}(\mathbf{g}_{DT}), & \mathbf{g}_{DM} \cdot \mathbf{g}_{DT} < 0, \\ \mathbf{g}_{DM} + \lambda \cdot \mathbf{g}_{DT}, & \text{else}, \end{cases} \tag{9}$$

where $\mathbf{g}_{DM}$ and $\mathbf{g}_{DT}$ denote the gradients ∇L_{DM} and ∇L_{DT}, respectively, and the hyper-parameter λ is employed to balance the downstream guidance (∇L_{DT}) and the diffusion loss (∇L_{DM}).

3 EXPERIMENTS

We evaluate our method and baseline approaches across four distinct meta-RL control tasks, following the dataset construction and experimental settings of Hu et al. [6]. Experimental results demonstrate that our approach significantly outperforms other parameter-efficient fine-tuning methods [14, 15], achieving performance levels close to the upper bound established by full-data fine-tuning. These findings highlight the distinct advantages of our proposed prompt-tuning technique in meta-RL settings.

4 DISCUSSION

Prompt Initialization. Unlike conventional prompt-tuning methods, which depend on high-quality prompts, Prompt Diffuser is robust to variations in training data and initialization. By modeling prompt-tuning as a conditional generative process, it refines prompts through downstream task guidance, enabling high-quality generation even from suboptimal initializations, and demonstrating its effectiveness without reliance on expert datasets.

Generative Models. Our approach's superiority stems from the diffusion model's ability to generate precise, in-distribution prompts, crucial for performance. While conditional generative modeling eliminates reliance on expert-curated prompts, it demands high precision. Leveraging the expressiveness of diffusion models, our method significantly outperforms other parameter-efficient approaches in prompt-tuning.

ACKNOWLEDGMENTS

This work was supported by STI 2030-Major Projects (No. 2021ZD 0201405), Shenzhen Basic Research Project (Natural Science Foundation) Basic Research Key Project (NO. JCYJ20241202124430041).

REFERENCES

[1] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* (2021).

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* (2020).

[3] Shengchao Hu, Ziqing Fan, Chaoqin Huang, Li Shen, Ya Zhang, Yanfeng Wang, and Dacheng Tao. 2024. Q-value Regularized Transformer for Offline Reinforcement Learning. *International Conference on Machine Learning* (2024).

[4] Shengchao Hu, Ziqing Fan, Li Shen, Ya Zhang, Yanfeng Wang, and Dacheng Tao. 2024. HarmoDT: Harmony Multi-Task Decision Transformer for Offline Reinforcement Learning. *International Conference on Machine Learning* (2024).

[5] Shengchao Hu, Li Shen, Ya Zhang, Yixin Chen, and Dacheng Tao. 2024. On Transforming Reinforcement Learning With Transformers: The Development Trajectory. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[6] Shengchao Hu, Li Shen, Ya Zhang, and Dacheng Tao. 2023. Prompt-Tuning Decision Transformer with Preference Ranking. *arXiv preprint arXiv:2305.09648* (2023).

[7] Shengchao Hu, Wanru Zhao, Weixiong Lin, Li Shen, Ya Zhang, and Dacheng Tao. 2024. Prompt Tuning with Diffusion for Few-Shot Pre-trained Policy Generalization. *arXiv preprint arXiv:2411.01168* (2024).

[8] Kuang-Huei Lee, Ofir Nachum, Mengjiao Sherry Yang, Lisa Lee, Daniel Freeman, Sergio Guadarrama, Ian Fischer, Winnie Xu, Eric Jang, Henryk Michalewski, et al. 2022. Multi-game decision transformers. *Advances in neural information processing systems* (2022).

[9] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[10] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. 2023. Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models. *arXiv preprint arXiv:2306.09251* (2023).

[11] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

[12] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning* (2015).

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

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

[15] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. 2022. Prompting decision transformer for few-shot policy generalization. *International Conference on Machine Learning* (2022).