Novelty Assessment Report

Paper: Diffusion Fine-Tuning via Reparameterized Policy Gradient of the Soft Q-Function

PDF URL: https://openreview.net/pdf?id=8zoxC9e23q

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-08

Abstract

Diffusion models excel at generating high-likelihood samples but often require alignment with downstream objectives. Existing fine-tuning methods for diffusion models significantly suffer from reward over-optimization, resulting in high-reward but unnatural samples and degraded diversity. To mitigate over-optimization, we propose Soft Q-based Diffusion Finetuning (SQDF), a novel KL-regularized RL method for diffusion alignment that applies a reparameterized policy gradient of a training-free, differentiable estimation of the soft Q-function. SQDF is further enhanced with three innovations: a discount factor for proper credit assignment in the denoising process, the integration of consistency models to refine Q-function estimates, and the use of an off-policy replay buffer to improve mode coverage and manage the reward-diversity trade-off. Our experiments demonstrate that SQDF achieves superior target rewards while preserving diversity in text-to-image alignment. Furthermore, in online black-box optimization, SQDF attains high sample efficiency while maintaining naturalness and diversity. Our code is available at https://anonymous.4open.science/r/SQDF-B66C

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Alignment of Diffusion Models with Downstream Objectives via Reinforcement Learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Fine-Tuning Methods and Algorithmic Frameworks**
- **Inference-Time Alignment and Guidance**
- **Survey and Tutorial Works**
- **Domain-Specific Applications**
- **Diffusion Models for Reinforcement Learning Tasks**
- **Related Alignment and Optimization Methods**

Complete Taxonomy Tree

- Alignment of Diffusion Models with Downstream Objectives via Reinforcement Learning Survey Taxonomy
- Fine-Tuning Methods and Algorithmic Frameworks
  - Policy Gradient and Direct RL Approaches
  - Foundational Policy Gradient Methods (3 papers)
    - [1] Training Diffusion Models with Reinforcement Learning (Black, 2023) View paper
    - [5] DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models (Ying, 2023) View paper
    - [10] Large-scale Reinforcement Learning for Diffusion Models (Yinan Zhang, 2024) View paper
  - Advanced Credit Assignment and Trajectory-Level Optimization (3 papers)
    - [14] Towards Better Alignment: Training Diffusion Models with Reinforcement Learning Against Sparse Rewards (Zijing Hu, 2025) View paper
    - [16] Aligning Few-Step Diffusion Models with Dense Reward Difference Learning (Zhang Ziyi, 2024) View paper
    - [18] Diffusion-Sharpening: Fine-tuning Diffusion Models with Denoising Trajectory Sharpening (Tian Ye, 2025) View paper
  - Continuous-Time and Stochastic Control Formulations (2 papers)
    - [20] Score as action: Fine-tuning diffusion generative models by continuous-time reinforcement learning (Zhao, 2025) View paper
    - [21] Scores as Actions: a framework of fine-tuning diffusion models by continuous-time reinforcement learning (Zhao, 2024) View paper
  - Soft Q-Function and Value-Based Methods ★ (2 papers)
    - [0] Diffusion Fine-Tuning via Reparameterized Policy Gradient of the Soft Q-Function (Anon et al., 2026) View paper
    - [29] Advantage Weighted Matching: Aligning RL with Pretraining in Diffusion Models (Ge, 2025) View paper
  - Preference-Based Alignment Methods (2 papers)
  - [8] Diffusion Model Alignment Using Direct Preference Optimization (Bram Wallace, 2024) View paper
  - [45] Divergence Minimization Preference Optimization for Diffusion Model Alignment (Li Binxu, 2025) View paper
  - Gradient-Based and Backpropagation Methods (2 papers)
  - [11] Aligning Text-to-Image Diffusion Models with Reward Backpropagation (Prabhudesai, 2023) View paper
  - [46] PRDP: Proximal Reward Difference Prediction for Large-Scale Reward Finetuning of Diffusion Models (Fei Deng, 2024) View paper
  - Specialized Alignment Objectives
  - Diversity and Bias Mitigation (2 papers)
    - [3] Training diffusion models towards diverse image generation with reinforcement learning (Zichen Miao, 2024) View paper

    - [24] Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback (He, 2024) View paper
  - Constrained and Regularized Alignment (1 papers)
    - [7] Aligning Text-to-Image Diffusion Models With Constrained Reinforcement Learning (Ziyi Zhang, 2025) View paper
  - Component-Specific Fine-Tuning (1 papers)
  - [9] Enhancing Diffusion Models with Text-Encoder Reinforcement Learning (Chao-Feng Chen, 2023) View paper
  - Scalability and Efficiency Innovations (2 papers)
  - [22] RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment (Dong, 2023) View paper
  - [23] DanceGRPO: Unleashing GRPO on Visual Generation (Xue, 2025) View paper
- Inference-Time Alignment and Guidance (3 papers)
  - [6] Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review (Uehara, 2025) View paper
  - [12] Inference-Time Alignment Control for Diffusion Models with Reinforcement Learning Guidance (Qiu Zijie, 2025) View paper
  - [13] Reward-guided controlled generation for inference-time alignment in diffusion models: Tutorial and review (Masatoshi Uehara, 2025) View paper
- Survey and Tutorial Works (3 papers)
  - [2] Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review (Uehara, 2024) View paper
  - [4] Enhancing Deep Reinforcement Learning: A Tutorial on Generative Diffusion Models in Network Optimization (Hongyang Du, 2024) View paper
  - [17] Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing (Wu Sihao, 2025) View paper
- Domain-Specific Applications
  - Image Editing and Inpainting (1 papers)
  - [34] PrefPaint: Aligning Image Inpainting Diffusion Model with Human Preference (Junhui Hou, 2024) View paper
  - 3D and Multi-View Generation (1 papers)
  - [35] Carve3D: Improving Multi-view Reconstruction Consistency for Diffusion Models with RL Finetuning (Desai Xie, 2024) View paper
  - Medical and Scientific Imaging (1 papers)
  - [39] RL4Med-DDPO: reinforcement learning for controlled guidance towards diverse medical image generation using vision-language foundation models (Parham Saremi, 2025) View paper
  - Distributed Systems and Service Optimization (1 papers)
  - [38] Reinforcement Learning With LLMs Interaction for Distributed Diffusion Model Services (Hongyang Du, 2025) View paper
- Diffusion Models for Reinforcement Learning Tasks
  - Diffusion Policies for Online and Offline RL (4 papers)
  - [19] Policy Representation via Diffusion Probability Model for Reinforcement Learning (Yang, 2023) View paper
  - [25] Aligning diffusion behaviors with q-functions for efficient continuous control (Huayu Chen, 2024) View paper
  - [37] Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning (Wang, 2022) View paper
  - [40] Efficient Diffusion Policies for Offline Reinforcement Learning (Kang, 2023) View paper
  - Diffusion-Based Planning and World Models (5 papers)
  - [26] BiTrajDiff: Bidirectional Trajectory Generation with Diffusion Models for Offline Reinforcement Learning (Qing, 2025) View paper
  - [44] Adaptdiffuser: Diffusion models as adaptive self-evolving planners (Liang Zhixuan, 2023) View paper
  - [47] Reasoning with latent diffusion in offline reinforcement learning (Venkatraman, 2023) View paper
  - [48] Diffusion-based offline rl for improved decision-making in augmented arc task (Kim Yunho, 2024) View paper
  - [49] MetaDiffuser: Diffusion Model as Conditional Planner for Offline Meta-RL (Ni Fei, 2023) View paper
  - Diffusion Policy Refinement with RL (4 papers)
  - [32] Maximize Your Diffusion: A Study into Reward Maximization and Alignment for Diffusion-based Control (Huh, 2025) View paper
  - [36] Diwa: Diffusion policy adaptation with world models (Nematollahi, 2025) View paper
  - [43] World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation (Jiang Zhennan, 2025) View paper
  - [50] Steering Your Diffusion Policy with Latent Space Reinforcement Learning (Wagenmaker, 2025) View paper
  - Specialized RL Applications with Diffusion (3 papers)
  - [27] Diffusion policy distillation for offline reinforcement learning. (Jia-zhi Zhang, 2025) View paper
  - [33] Principled and Tractable RL for Reasoning with Diffusion Language Models (Zhan, 2025) View paper
  - [41] Toward multi-task generalization in autonomous navigation: A human-in-the-loop adversarial reinforcement learning with diffusion policy (Dong Hu, 2025) View paper
- Related Alignment and Optimization Methods
  - Non-Diffusion Generative Model Alignment (2 papers)
  - [15] AlignSAM: Aligning Segment Anything Model to Open Context via Reinforcement Learning (Duojun Huang, 2024) View paper
  - [28] Prompt Optimizer of Text-to-Image Diffusion Models for Abstract Concept Understanding (Li, 2024) View paper
  - Transferable and Meta-Learning Methods (3 papers)
  - [30] Transferable reinforcement learning via generalized occupancy models (Chuning Zhu, 2024) View paper
  - [31] Adding conditional control to diffusion models with reinforcement learning (Zhao, 2024) View paper
  - [42] Reinforcement learning with formation energy feedback for material diffusion models. (Jiao Huang, 2025) View paper

Narrative

Core task: alignment of diffusion models with downstream objectives via reinforcement learning. The field organizes around several major branches that reflect different strategic emphases. Fine-tuning methods and algorithmic frameworks encompass policy gradient and value-based techniques that directly optimize diffusion model parameters, with works like Training Diffusion RL[1] and DPOK[5] exemplifying direct RL approaches. Inference-time alignment and guidance methods, such as Inference Time Alignment[6] and Inference Time Control[12], adjust generation without retraining by steering the sampling process. Domain-specific applications demonstrate how these alignment strategies translate to robotics, medical imaging, and other specialized settings, while a growing body of survey and tutorial works, including RL Finetuning Tutorial[2] and Preference Alignment Survey[17], synthesizes emerging best practices. Related alignment and optimization methods explore connections to preference learning and divergence minimization, and a separate branch addresses diffusion models for reinforcement learning tasks, where generative models serve as policy representations or world models.

Within fine-tuning frameworks, a particularly active line of research contrasts policy gradient methods with value-based and soft Q-function approaches. Policy gradient techniques often face challenges with high variance and credit assignment across diffusion timesteps, prompting explorations of advantage weighting and variance reduction. Value-based methods, including soft Q-function formulations, offer an alternative by learning action-value estimates to guide alignment. Reparameterized Policy Gradient[0] sits squarely in this value-based cluster, emphasizing soft Q-function techniques to stabilize gradient estimation. It shares methodological kinship with Advantage Weighted Matching[29], which also leverages advantage estimates to refine diffusion policies, yet differs in how it reparameterizes the optimization objective. These approaches collectively address the tension between sample efficiency and stability, a central open question as practitioners scale alignment to large models and diverse reward signals.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Advantage Weighted Matching: Aligning RL with Pretraining in Diffusion Models

Authors: Ge, Chongjian, Shuchen Xue, Zhang Shilong, Chongjian Ge, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement Learning (RL) has emerged as a central paradigm for advancing Large Language Models (LLMs), where pre-training and RL post-training share the same log-likelihood formulation. In contrast, recent RL approaches for diffusion models, most notably Denoising Diffusion Policy Optimization (DDPO), optimize an objective different from the pretraining objectives--score/flow matching loss. In this work, we establish a novel theoretical analysis: DDPO is an implicit form of score/flow matchin...

Relationship Analysis

Both papers belong to the Soft Q-Function and Value-Based Methods category, employing KL-regularized RL frameworks for diffusion model alignment. They share the goal of mitigating reward over-optimization through soft Q-function approaches and KL regularization with pretrained models. However, the original paper (SQDF) uses a training-free soft Q-function approximation via posterior mean estimation with consistency models and introduces a discount factor for credit assignment, while the candidate paper (AWM) reformulates DDPO as denoising score matching with noisy data and proposes advantage-weighted matching that directly applies the pretraining score/flow matching objective with reward-based sample reweighting.

Contributions Analysis

Overall novelty summary. The paper proposes SQDF, a KL-regularized RL method using soft Q-function estimation for diffusion model alignment. It resides in the 'Soft Q-Function and Value-Based Methods' leaf, which contains only two papers total (including this work). This leaf sits within the broader 'Policy Gradient and Direct RL Approaches' branch, which encompasses foundational policy gradient methods, advanced credit assignment techniques, and continuous-time formulations. The sparse population of this specific leaf suggests that value-based approaches remain relatively underexplored compared to policy gradient methods in diffusion alignment.

The taxonomy reveals that most fine-tuning work concentrates in adjacent leaves: 'Foundational Policy Gradient Methods' (3 papers), 'Advanced Credit Assignment and Trajectory-Level Optimization' (3 papers), and 'Continuous-Time and Stochastic Control Formulations' (2 papers). The sibling paper in this leaf, Advantage Weighted Matching, also leverages advantage estimates but differs in optimization formulation. Neighboring branches include preference-based alignment methods and gradient-based backpropagation approaches, which avoid explicit RL formulations. The taxonomy's scope and exclude notes clarify that this leaf focuses specifically on value function estimation, distinguishing it from pure policy gradient techniques.

Among nine candidates examined, the core SQDF method (Contribution A) shows one refutable candidate from seven examined, suggesting some prior overlap in soft Q-function approaches for diffusion alignment. The three stabilization techniques (Contribution B) examined two candidates with no refutations, indicating these enhancements may be more novel. The training-free Q-function approximation (Contribution C) examined zero candidates, leaving its novelty unassessed within this limited search. The statistics reflect a focused semantic search rather than exhaustive coverage, so contributions without clear refutations may still have undiscovered prior work.

Given the limited search scope (nine candidates total), this analysis captures immediate semantic neighbors but cannot rule out relevant work outside the top-K matches. The sparse leaf population and single refutable pair suggest SQDF occupies a relatively less-crowded methodological niche, though the field's rapid growth means recent preprints or concurrent work may not appear in this snapshot. The enhancement techniques appear more distinctive within the examined literature than the core soft Q-function framework.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Soft Q-based Diffusion Finetuning (SQDF) method

Description: The authors introduce SQDF, a KL-regularized reinforcement learning framework that fine-tunes diffusion models using a reparameterized policy gradient guided by a training-free soft Q-function approximation. This approach avoids unstable value function training and enables direct use of reward gradients for low-variance policy updates while mitigating reward over-optimization.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Value Diffusion Reinforcement Learning

URL: View paper

Brief Assessment

Value Diffusion RL[57] focuses on using diffusion models to represent value distributions in distributional RL for improved Q-value estimation, not on KL-regularized policy gradient methods for diffusion model alignment. The candidate addresses value function learning rather than policy fine-tuning with soft Q-functions.

2. Forward kl regularized preference optimization for aligning diffusion policies

URL: View paper

Brief Assessment

Forward KL Regularized[54] focuses on aligning diffusion policies with human preferences using direct preference optimization (DPO) in a two-stage framework, whereas SQDF addresses reward optimization through KL-regularized RL with a training-free soft Q-function approximation for policy gradient updates. The technical approaches and problem formulations differ fundamentally.

3. Amortizing intractable inference in diffusion models for vision, language, and control

URL: View paper

Brief Assessment

Amortizing Intractable Inference[53] focuses on posterior inference under diffusion priors using trajectory balance objectives for various tasks (vision, language, control), while SQDF specifically addresses KL-regularized RL for diffusion alignment using soft Q-function approximation with discount factors and consistency models. The technical approaches and problem formulations differ substantially.

4. PADRE: Pseudo-likelihood based alignment of diffusion language models

URL: View paper

Brief Assessment

PADRE[55] focuses on alignment of diffusion language models using pseudo-likelihood objectives, while the original paper addresses KL-regularized RL for general diffusion models (e.g., text-to-image) using soft Q-function approximations. The technical domains and model types differ fundamentally.

5. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models

URL: View paper

Prior Art Analysis

DPOK[5] demonstrates that KL-regularized reinforcement learning for diffusion model alignment using policy gradients was previously proposed and implemented. DPOK[5] presents a method that fine-tunes diffusion models using policy gradient with KL regularization, incorporating a reward function and KL divergence terms in the objective. The candidate paper shows this approach was published at NeurIPS 2023, predating the original paper's submission. Both papers frame diffusion fine-tuning as an MDP, use policy gradient methods with KL regularization to prevent over-optimization, and directly leverage reward gradients for model updates.

Evidence

Evidence 1 - **Rationale**: Both papers propose KL-regularized RL methods for fine-tuning diffusion models using policy gradient approaches. DPOK[5] was published at NeurIPS 2023, establishing prior work in this area. - **Original**: we proposesoft q-based diffusion finetuning (sqdf), a novel kl-regularized rl method for diffusion alignment that applies a reparameterized policy gradient of a training-free, differentiable estimation of the soft q-function. - **Candidate**: we propose using online reinforcement learning (rl) to fine-tune text-to-image models. we focus on diffusion models, defining the fine-tuning task as an rl problem, and updating the pre-trained text-to-image diffusion models using policy gradient to maximize the feedbacktrained reward. our approach,...

Evidence 2 - **Rationale**: Both papers use KL regularization with the pre-trained model to prevent over-optimization during fine-tuning, demonstrating that this approach was already established in DPOK[5]. - **Original**: to this end, we proposesoft q-based diffusion finetuning (sqdf), a method that employs a reparameterized policy gradient guided by a training-free soft q-function within a kl-regularized rl framework. - **Candidate**: adding kl regularization. the risk of fine-tuning purely based on the reward model learned from human or ai feedback is that the model may overfit to the reward and discount the "skill" of the initial diffusion model to a greater degree than warranted. to avoid this phenomenon, similar to [23, 41], ...

Evidence 3 - **Rationale**: Both papers use policy gradient methods to update diffusion models by leveraging reward gradients without backpropagating through the entire denoising trajectory, showing DPOK[5] established this approach first. - **Original**: sqdf leverages a training-free, one-step soft q-function approximation to directly apply the reward gradient via a reparameterized policy update, effectively fine-tuning the model without backpropagation through the denoising chain. - **Candidate**: equation (6) is equivalent to the gradient used by the popular policy gradient algorithm, reinforce, to update a policy in the mdp (4). the gradient in (6) is estimated from trajectories pθpx0:t |zq generated by the current policy, and then used to update the policypθpxt´1|xt, zqin an online fashion...

Evidence 4 - **Rationale**: DPOK[5] provides the mathematical framework for computing policy gradients in diffusion models that enable direct use of reward gradients, establishing the foundation for this approach. - **Original**: the core of our approach is to approximate the soft q-function via a single-step posterior mean approximation (li et al., 2024), a strategy that circumvents the need for unstable value function learning. this approximation is differentiable under the parameterized oracle or proxy models (xu et al., ... - **Candidate**: lemma 4.1 (a modification of theorem 4.1 in [5]). if pθpx0:t |zqrpx0, zqand ∇θpθpx0:t |zqrpx0, zq are continuous functions ofθ, then we can write the gradient of the objective in(5) as ∇θeppzqepθpx0|zqr´rpx0, zqs" eppzqepθpx0:t |zq « ´rpx0, zq tÿ t"1 ∇θ log pθpxt´1|xt, zq ff .

6. Residual Policy Gradient: A Reward View of KL-regularized Objective

URL: View paper

Brief Assessment

Residual Policy Gradient[56] focuses on policy gradient methods for general RL and policy customization in continuous control tasks (MuJoCo), not on diffusion model fine-tuning or alignment with reward functions through soft Q-function approximations.

7. Controllable Diffusion via Optimal Classifier Guidance

URL: View paper

Brief Assessment

Optimal Classifier Guidance[58] focuses on supervised learning-based controllable diffusion using iterative classifier training, not KL-regularized RL with soft Q-function policy gradients. The candidate's approach trains a lightweight classifier to guide generation without modifying the base model, whereas SQDF employs reparameterized policy gradients with training-free soft Q-function approximation for direct diffusion model fine-tuning.

Contribution 2: Three stabilization and enhancement techniques for SQDF

Description: The authors propose three complementary techniques to improve SQDF: (1) a discount factor gamma to downweight early denoising steps for better credit assignment, (2) consistency models to provide more accurate soft Q-function estimates across all timesteps, and (3) an off-policy replay buffer to improve mode coverage and manage the reward-diversity trade-off.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Regularized conditional diffusion model for multi-task preference alignment

URL: View paper

Brief Assessment

Multi Task Preference[51] focuses on multi-task preference alignment using representation learning and mutual information regularization for diffusion models, not on stabilizing Q-function estimation in single-task RL with discount factors, consistency models, and replay buffers as proposed in the original paper.

2. Distributed and Controllable Mobile Text-to-Image Generation with User Preference Guarantee

URL: View paper

Brief Assessment

Distributed Mobile Generation[52] focuses on distributed text-to-image generation on mobile devices with user preference optimization, not on stabilizing diffusion model fine-tuning through discount factors, consistency models, and replay buffers for credit assignment and Q-function estimation as in the original paper.

Contribution 3: Training-free soft Q-function approximation via posterior mean

Description: The authors develop a training-free approximation of the soft Q-function using single-step posterior mean estimation derived from Tweedie's formula. This approximation is differentiable under parameterized reward models, enabling direct gradient-based policy updates without requiring explicit value function training.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Diffusion Fine-Tuning via Reparameterized Policy Gradient of the Soft Q-Function View paper
- [1] Training Diffusion Models with Reinforcement Learning View paper
- [2] Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review View paper
- [3] Training diffusion models towards diverse image generation with reinforcement learning View paper
- [4] Enhancing Deep Reinforcement Learning: A Tutorial on Generative Diffusion Models in Network Optimization View paper
- [5] DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models View paper
- [6] Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review View paper
- [7] Aligning Text-to-Image Diffusion Models With Constrained Reinforcement Learning View paper
- [8] Diffusion Model Alignment Using Direct Preference Optimization View paper
- [9] Enhancing Diffusion Models with Text-Encoder Reinforcement Learning View paper
- [10] Large-scale Reinforcement Learning for Diffusion Models View paper
- [11] Aligning Text-to-Image Diffusion Models with Reward Backpropagation View paper
- [12] Inference-Time Alignment Control for Diffusion Models with Reinforcement Learning Guidance View paper
- [13] Reward-guided controlled generation for inference-time alignment in diffusion models: Tutorial and review View paper
- [14] Towards Better Alignment: Training Diffusion Models with Reinforcement Learning Against Sparse Rewards View paper
- [15] AlignSAM: Aligning Segment Anything Model to Open Context via Reinforcement Learning View paper
- [16] Aligning Few-Step Diffusion Models with Dense Reward Difference Learning View paper
- [17] Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing View paper
- [18] Diffusion-Sharpening: Fine-tuning Diffusion Models with Denoising Trajectory Sharpening View paper
- [19] Policy Representation via Diffusion Probability Model for Reinforcement Learning View paper
- [20] Score as action: Fine-tuning diffusion generative models by continuous-time reinforcement learning View paper
- [21] Scores as Actions: a framework of fine-tuning diffusion models by continuous-time reinforcement learning View paper
- [22] RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment View paper
- [23] DanceGRPO: Unleashing GRPO on Visual Generation View paper
- [24] Aligning as Debiasing: Causality-Aware Alignment via Reinforcement Learning with Interventional Feedback View paper
- [25] Aligning diffusion behaviors with q-functions for efficient continuous control View paper
- [26] BiTrajDiff: Bidirectional Trajectory Generation with Diffusion Models for Offline Reinforcement Learning View paper
- [27] Diffusion policy distillation for offline reinforcement learning. View paper
- [28] Prompt Optimizer of Text-to-Image Diffusion Models for Abstract Concept Understanding View paper
- [29] Advantage Weighted Matching: Aligning RL with Pretraining in Diffusion Models View paper
- [30] Transferable reinforcement learning via generalized occupancy models View paper
- [31] Adding conditional control to diffusion models with reinforcement learning View paper
- [32] Maximize Your Diffusion: A Study into Reward Maximization and Alignment for Diffusion-based Control View paper
- [33] Principled and Tractable RL for Reasoning with Diffusion Language Models View paper
- [34] PrefPaint: Aligning Image Inpainting Diffusion Model with Human Preference View paper
- [35] Carve3D: Improving Multi-view Reconstruction Consistency for Diffusion Models with RL Finetuning View paper
- [36] Diwa: Diffusion policy adaptation with world models View paper
- [37] Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning View paper
- [38] Reinforcement Learning With LLMs Interaction for Distributed Diffusion Model Services View paper
- [39] RL4Med-DDPO: reinforcement learning for controlled guidance towards diverse medical image generation using vision-language foundation models View paper
- [40] Efficient Diffusion Policies for Offline Reinforcement Learning View paper
- [41] Toward multi-task generalization in autonomous navigation: A human-in-the-loop adversarial reinforcement learning with diffusion policy View paper
- [42] Reinforcement learning with formation energy feedback for material diffusion models. View paper
- [43] World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation View paper
- [44] Adaptdiffuser: Diffusion models as adaptive self-evolving planners View paper
- [45] Divergence Minimization Preference Optimization for Diffusion Model Alignment View paper
- [46] PRDP: Proximal Reward Difference Prediction for Large-Scale Reward Finetuning of Diffusion Models View paper
- [47] Reasoning with latent diffusion in offline reinforcement learning View paper
- [48] Diffusion-based offline rl for improved decision-making in augmented arc task View paper
- [49] MetaDiffuser: Diffusion Model as Conditional Planner for Offline Meta-RL View paper
- [50] Steering Your Diffusion Policy with Latent Space Reinforcement Learning View paper
- [51] Regularized conditional diffusion model for multi-task preference alignment View paper
- [52] Distributed and Controllable Mobile Text-to-Image Generation with User Preference Guarantee View paper
- [53] Amortizing intractable inference in diffusion models for vision, language, and control View paper
- [54] Forward kl regularized preference optimization for aligning diffusion policies View paper
- [55] PADRE: Pseudo-likelihood based alignment of diffusion language models View paper
- [56] Residual Policy Gradient: A Reward View of KL-regularized Objective View paper
- [57] Value Diffusion Reinforcement Learning View paper
- [58] Controllable Diffusion via Optimal Classifier Guidance View paper