Novelty Assessment Report

Paper: wd1: Weighted Policy Optimization for Reasoning in Diffusion Language Models
PDF URL: https://openreview.net/pdf?id=L2rfd2Czbj
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Improving the reasoning capabilities of diffusion-based large language models (dLLMs) through reinforcement learning (RL) remains an open problem. The intractability of dLLMs likelihood function necessitates approximating the current, old, and reference policy likelihoods at each policy optimization step. This reliance introduces additional computational overhead, and can lead to large variance and estimation error in RL objective -- particularly in computing the policy ratio for importance sampling. To mitigate these issues, we introduce wd1, a novel ratio-free policy optimization approach that reformulates the objective as a weighted log-likelihood, requiring only a single approximation for the current parametrized policy likelihood. We formally show that our proposed method can be interpreted as energy-guided discrete diffusion training combined with negative sample unlearning, thereby confirming its theoretical soundness. In experiments on LLaDA-8B model, \textit{wd1} outperforms diffusion-based GRPO (\textit{d1}) while requiring lower computational cost, achieving up to a $+59\%$ improvement in accuracy. Furthermore, we extend \textit{wd1} to denoising-stepwise weighted policy optimization (\algname++), achieving state-of-the-art math performance of 44.2% on MATH500 and 84.5% on GSM8K with only 20 RL training steps.

Core Task Landscape

This paper addresses: **Reinforcement Learning for Diffusion-Based Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **26 categories**.

Taxonomy Overview

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

- **Policy Optimization Algorithms for Diffusion Language Models**
- **Exploration and Inference Strategies for Diffusion Language Models**
- **Multimodal and Unified Diffusion Architectures**
- **Diffusion Models for Reinforcement Learning Tasks**
- **Reinforcement Learning for General Diffusion Model Alignment**
- **Application-Specific Diffusion and RL Integration**
- **Consolidation and Multimodal RL for Discrete Diffusion**
- **Surveys, Benchmarks, and Comparative Studies**
- **Open-Source Diffusion Language Model Implementations**

Complete Taxonomy Tree

- Reinforcement Learning for Diffusion-Based Large Language Models Survey Taxonomy
- Policy Optimization Algorithms for Diffusion Language Models
  - Likelihood Approximation and Variance Reduction ★ (5 papers)
  - [0] wd1: Weighted Policy Optimization for Reasoning in Diffusion Language Models (Anon et al., 2026) View paper
  - [8] LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models (Zhu Feng-qi, 2025) View paper
  - [14] Boundary-Guided Policy Optimization for Memory-efficient RL of Diffusion Large Language Models (Lin, 2025) View paper
  - [16] SPG: Sandwiched Policy Gradient for Masked Diffusion Language Models (Wang Chen-yu, 2025) View paper
  - [37] d2: Improved Techniques for Training Reasoning Diffusion Language Models (Wang Guanghan, 2025) View paper
  - Trajectory-Aware and Step-Aware Policy Optimization (3 papers)
  - [9] Step-Aware Policy Optimization for Reasoning in Diffusion Large Language Models (Xie Shao-an, 2025) View paper
  - [11] Revolutionizing reinforcement learning framework for diffusion large language models (Wang Yinjie, 2025) View paper
  - [12] Principled and Tractable RL for Reasoning with Diffusion Language Models (Zhan, 2025) View paper
  - Distribution Matching and Alignment Methods (3 papers)
  - [22] Enhancing reasoning for diffusion llms via distribution matching policy optimization (Zhu Yuchen, 2025) View paper
  - [24] Improving Reasoning for Diffusion Language Models via Group Diffusion Policy Optimization (Lin Jia-He, 2025) View paper
  - [25] DiFFPO: Training Diffusion LLMs to Reason Fast and Furious via Reinforcement Learning (Zhao, 2025) View paper
  - Reliable and Tree-Structured Policy Optimization (1 papers)
  - [44] d-TreeRPO: Towards More Reliable Policy Optimization for Diffusion Language Models (Leyi Pan, 2025) View paper
- Exploration and Inference Strategies for Diffusion Language Models
  - Inpainting-Guided and Exploration-Enhanced Training (2 papers)
  - [4] Reinforcing the diffusion chain of lateral thought with diffusion language models (Huang Ze-min, 2025) View paper
  - [7] Inpainting-Guided Policy Optimization for Diffusion Large Language Models (Zhao Siyan, 2025) View paper
  - Unmasking Order and Decoding Strategy Optimization (2 papers)
  - [27] Learning Unmasking Policies for Diffusion Language Models (Metod Jazbec, 2025) View paper
  - [38] Lookahead Unmasking Elicits Accurate Decoding in Diffusion Language Models (Sanghyun Lee, 2025) View paper

- Self-Reflective and Iterative Refinement (1 papers)
  - [40] Don't Settle Too Early: Self-Reflective Remasking for Diffusion Language Models (Huang Ze-min, 2025) View paper
  - Training-Inference Discrepancy Resolution (1 papers)
  - [2] Mdpo: Overcoming the training-inference divide of masked diffusion language models (He, 2025) View paper
- Multimodal and Unified Diffusion Architectures
  - Multimodal Diffusion Language Models (2 papers)
  - [1] MMaDA: Multimodal Large Diffusion Language Models (Yang Ling, 2025) View paper
  - [31] MMaDA-Parallel: Multimodal Large Diffusion Language Models for Thinking-Aware Editing and Generation (Ye Tian, 2025) View paper
  - Unified Reinforcement Learning Frameworks (1 papers)
  - [39] UniRL-Zero: Reinforcement Learning on Unified Models with Joint Language Model and Diffusion Model Experts (Wang, 2025) View paper
- Diffusion Models for Reinforcement Learning Tasks
  - Diffusion-Based Policy Representation and Training (3 papers)
  - [32] Diffusion-based Reinforcement Learning via Q-weighted Variational Policy Optimization (Ding, 2024) View paper
  - [49] Modular Diffusion Policy Training: Decoupling and Recombining Guidance and Diffusion for Offline RL (Chen Zhaoyang, 2025) View paper
  - [50] TARAD: Task-Aware Robot Affordance-Centric Diffusion Policy Learned From LLM-Generated Demonstrations (Site Hu, 2025) View paper
  - Trajectory Stitching and Data Augmentation (1 papers)
  - [19] DiffStitch: Boosting Offline Reinforcement Learning with Diffusion-based Trajectory Stitching (Li Guanghe, 2024) View paper
  - Multi-Task Planning and Generalist Agents (2 papers)
  - [30] Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning (He, 2023) View paper
  - [34] Diffusion augmented agents: A framework for efficient exploration and transfer learning (Di Palo, 2024) View paper
  - Scene Reconstruction and Simulation Enhancement (1 papers)
  - [43] ReconDreamer-RL: Enhancing Reinforcement Learning via Diffusion-based Scene Reconstruction (Ni Chaojun, 2025) View paper
- Reinforcement Learning for General Diffusion Model Alignment
  - Denoising Diffusion Policy Optimization (2 papers)
  - [5] Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review (Uehara, 2024) View paper
  - [18] Training Diffusion Models with Reinforcement Learning (Black, 2023) View paper
  - Inference-Time Alignment and Control (1 papers)
  - [23] Inference-Time Alignment Control for Diffusion Models with Reinforcement Learning Guidance (Qiu Zijie, 2025) View paper
  - Value-Based and Dense Reward Methods (1 papers)
  - [17] VARD: Efficient and Dense Fine-Tuning for Diffusion Models with Value-based RL (Zhuang, 2025) View paper
  - Amortized Inference and Posterior Sampling (1 papers)
  - [13] Amortizing intractable inference in diffusion models for vision, language, and control (Venkatraman, 2024) View paper
- Application-Specific Diffusion and RL Integration
  - Image Generation and Visual Reasoning (3 papers)
  - [26] Reasoning physical video generation with diffusion timestep tokens via reinforcement learning (Lin Wang, 2025) View paper
  - [36] Learning profitable NFT image diffusions via multiple visual-policy guided reinforcement learning (Huiguo He, 2023) View paper
  - [42] Self-Reflective Reinforcement Learning for Diffusion-based Image Reasoning Generation (Pan Jiadong, 2025) View paper
  - Distributed Systems and Quality of Service Optimization (3 papers)
  - [10] Reinforcement Learning With LLMs Interaction for Distributed Diffusion Model Services (Hongyang Du, 2023) View paper
  - [20] User-centric interactive AI for distributed diffusion model-based AI-generated content (DU Hongyang, 2023) View paper
  - [21] Enhancing LLM QoS through Cloud-Edge Collaboration: A Diffusion-based Multi-Agent Reinforcement Learning Approach (Zhi Yao, 2025) View paper
  - Empathetic and Multimodal Response Generation (2 papers)
  - [35] ReflectDiffu:Reflect between Emotion-intent Contagion and Mimicry for Empathetic Response Generation via a RL-Diffusion Framework (Yuan, 2024) View paper
  - [48] Make Imagination Clearer! Stable Diffusion-based Visual Imagination for Multimodal Machine Translation (Andong Chen, 2025) View paper
  - Tool Orchestration and Interleaved Generation (1 papers)
  - [41] LLM-I: LLMs are Naturally Interleaved Multimodal Creators (Zhang Feng, 2025) View paper
  - Specialized Domain Applications (2 papers)
  - [29] Hybrid RAG-Empowered Multimodal LLM for Secure Data Management in Internet of Medical Things: A Diffusion-Based Contract Approach (Cheng Su, 2024) View paper
  - [33] LLM-Assisted Red Teaming of Diffusion Models through" Failures Are Fated, But Can Be Faded" (Som Sagar, 2024) View paper
- Consolidation and Multimodal RL for Discrete Diffusion (1 papers)
  - [6] Consolidating Reinforcement Learning for Multimodal Discrete Diffusion Models (Ma Tianren, 2025) View paper
- Surveys, Benchmarks, and Comparative Studies (3 papers)
  - [3] Diffusion-based Large Language Models Survey (Chiung-Yi Tseng, 2025) View paper
  - [28] Unifying Modern AI with Robotics: Survey on MDPs with Diffusion and Foundation Models (Zhaofan Zhang, 2025) View paper
  - [45] Comparative Evaluation of Reasoning and Inference in LLM-Based and Diffusion-Based Approaches (W Zhao, 2025) View paper
- Open-Source Diffusion Language Model Implementations (3 papers)
  - [15] d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning (Zhao Siyan, 2025) View paper
  - [46] d1: Scaling Reasoning in Dii¬usion Large Language Models via Reinforcement Learning (Zhao Siyan, 2025) View paper
  - [47] Dream-Coder 7B: An Open Diffusion Language Model for Code (Xie, 2025) View paper

Narrative

Core task: Reinforcement learning for diffusion-based large language models. The field has rapidly expanded into several major branches that reflect different emphases in adapting RL to discrete diffusion architectures. Policy Optimization Algorithms for Diffusion Language Models focuses on developing tractable gradient estimators and variance reduction techniques—such as likelihood approximation

methods—that enable stable training of diffusion-based text generators under reward signals. Exploration and Inference Strategies examine how to guide the iterative denoising process at test time, while Multimodal and Unified Diffusion Architectures extend these ideas beyond text to vision and cross-modal settings. Meanwhile, Diffusion Models for Reinforcement Learning Tasks and Application-Specific Diffusion and RL Integration explore using diffusion as a planning or policy representation in traditional RL domains, and Reinforcement Learning for General Diffusion Model Alignment addresses broader safety and preference alignment questions. Surveys, Benchmarks, and Comparative Studies provide overarching perspectives, and Open-Source Diffusion Language Model Implementations offer practical tooling for the community.

Within Policy Optimization, a dense cluster of works tackles the challenge of high-variance gradients inherent in discrete diffusion. LLaDA Variance Reduced[8] and Sandwiched Policy Gradient[16] exemplify efforts to bound or reduce variance through control variates and tighter likelihood bounds, while Boundary-Guided Policy[14] and d2 Improved Techniques[37] propose alternative parameterizations or training schedules. Weighted Policy Optimization[0] sits squarely in this Likelihood Approximation and Variance Reduction subgroup, sharing the goal of making policy gradients more stable and sample-efficient. Compared to LLaDA Variance Reduced[8], which emphasizes amortized baselines, Weighted Policy Optimization[0] explores weighting schemes that directly modulate gradient contributions across diffusion steps. This contrasts with Sandwiched Policy Gradient[16], which instead sandwiches the policy between upper and lower likelihood bounds. Across these neighboring works, the central trade-off remains balancing computational overhead against variance reduction, with each method offering a distinct lens on tractable credit assignment in diffusion language models.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models

Authors: Zhu Feng-qi, Wang Rong-zhen, Fengqi Zhu, Nie Shen, Rongzheng Wang, et al. (25 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

While Masked Diffusion Models (MDMs), such as LLaDA, present a promising paradigm for language modeling, there has been relatively little effort in aligning these models with human preferences via reinforcement learning. The challenge primarily arises from the high variance in Evidence Lower Bound (ELBO)-based likelihood estimates required for preference optimization. To address this issue, we propose Variance-Reduced Preference Optimization (VRPO), a framework that formally analyzes the varianc...

Relationship Analysis

Both papers belong to the Likelihood Approximation and Variance Reduction category, addressing the intractable likelihood problem in diffusion language models through improved policy gradient estimation. They overlap in their focus on reducing variance in ELBO-based likelihood estimates for policy optimization in masked diffusion models like LLaDA. The key difference is that the original paper (wd1) proposes a ratio-free weighted log-likelihood objective that eliminates policy ratio computation entirely, while the candidate paper (LLaDA 1.5/VRPO) focuses on variance reduction techniques like optimal Monte Carlo budget allocation and antithetic sampling while still computing likelihood ratios within a preference optimization framework.

---

2. Boundary-Guided Policy Optimization for Memory-efficient RL of Diffusion Large Language Models

Authors: Lin, Nianyi, Zhang Jiajie, Nianyi Lin, Hou Lei, et al. (10 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

A key challenge in applying reinforcement learning (RL) to diffusion large language models (dLLMs) lies in the intractability of their likelihood functions, which are essential for the RL objective, necessitating corresponding approximation in each training step. While existing methods approximate the log-likelihoods by their evidence lower bounds (ELBOs) via customized Monte Carlo (MC) sampling, the forward computational graphs of all MC samples need to be retained for the gradient computation ...

Relationship Analysis

Both papers belong to the Likelihood Approximation and Variance Reduction category, addressing the intractable likelihood problem in diffusion-based LLMs through improved policy gradient methods. While the original paper (wd1) eliminates policy ratios entirely by reformulating the RL objective as a weighted log-likelihood requiring only single likelihood approximation, the candidate paper (BGPO) focuses on memory-efficient Monte Carlo sampling by constructing a linear lower bound of the ELBO-based objective that enables gradient accumulation. The key difference is that wd1 avoids ratio computation to reduce variance and bias, whereas BGPO maintains ELBO-based approximation but optimizes memory usage to allow larger sample sizes for more accurate likelihood estimation.

---

3. SPG: Sandwiched Policy Gradient for Masked Diffusion Language Models

Authors: Wang Chen-yu, Rashidinejad, Paria, Chengyu Wang, Su, et al. (29 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Diffusion large language models (dLLMs) are emerging as an efficient alternative to autoregressive models due to their ability to decode multiple tokens in parallel. However, aligning dLLMs with human preferences or task-specific rewards via reinforcement learning (RL) is challenging because their intractable log-likelihood precludes the direct application of standard policy gradient methods. While prior work uses surrogates like the evidence lower bound (ELBO), these one-sided approximations ca...

Relationship Analysis

Both papers address the likelihood approximation problem in reinforcement learning for masked diffusion language models, sharing the same taxonomy category of developing improved estimators for policy gradients. While wd1 proposes a ratio-free weighted log-likelihood approach that eliminates policy ratio computation entirely, SPG introduces a sandwiched estimator using both upper and lower bounds of the log-likelihood to reduce bias in policy gradient estimation. The key difference is that wd1 reformulates the RL objective to avoid ratios altogether, whereas SPG improves the accuracy of likelihood-based policy gradient computation through tighter bounds.

---

4. d2: Improved Techniques for Training Reasoning Diffusion Language Models

Authors: Wang Guanghan, Schiff, Yair, Guanghan Wang, Yair Schiff, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

While diffusion language models (DLMs) have achieved competitive performance in text generation, improving their reasoning ability with reinforcement learning remains an active research area. Here, we introduce d2, a reasoning framework tailored for masked DLMs. Central to our framework is a new policy gradient algorithm that relies on properties of masking to accurately estimate the likelihoods of sampling trajectories. Our estimators trade off computation for approximation accuracy in an analy...

Relationship Analysis

Both papers belong to the Likelihood Approximation and Variance Reduction category, addressing the intractable likelihood problem in diffusion language models through improved policy gradient methods. They overlap in reformulating RL objectives to reduce computational overhead and variance in policy ratio estimation, with both proposing alternatives to standard diffusion-based GRPO. The key difference is that the original paper (wd1) eliminates policy ratios entirely through a weighted log-likelihood objective with advantage-based weights, while the candidate paper (d2) retains importance sampling but introduces novel trajectory likelihood estimators (StepMerge and AnyOrder) that reduce the number of forward passes required for accurate likelihood computation.

Contributions Analysis

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

Contribution 1: wd1: Weighted Policy Optimization for Diffusion Language Models

Description: The authors propose wd1, a reinforcement learning method for diffusion-based large language models that eliminates the need for policy ratio computation in importance sampling. The method reformulates the RL objective as a weighted log-likelihood where weights balance increasing probability of high-advantage completions and decreasing probability of low-advantage ones, requiring only one likelihood approximation instead of three.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Advantage weighted matching: Aligning rl with pretraining in diffusion models

URL: View paper

Brief Assessment

Advantage Weighted Matching[51] focuses on continuous diffusion models for image generation, not discrete diffusion language models. The technical domains and model architectures are fundamentally different.

Contribution 2: Theoretical interpretation as energy-guided diffusion with unlearning

Description: The authors establish a theoretical connection showing that their weighted policy optimization objective is equivalent to training an energy-guided discrete diffusion model where the energy function is the negative advantage, combined with unlearning of low-advantage samples. This provides formal justification for the method's design.

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.

Contribution 3: wd1++: Denoising-stepwise weighted policy optimization

Description: The authors extend their base method to wd1++, which leverages intermediate completions generated during the iterative denoising process rather than only using final outputs. This extension achieves state-of-the-art performance on mathematical reasoning benchmarks with fewer training steps and rollouts.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Efficient Online Reinforcement Learning for Diffusion Policy

URL: View paper

Brief Assessment

Efficient Online RL[54] focuses on online RL for continuous control with diffusion policies in robotics tasks, not on leveraging intermediate completions in discrete diffusion language models for reasoning tasks.

2. Training Diffusion Models with Reinforcement Learning

URL: View paper

Brief Assessment

Training Diffusion RL[18] focuses on text-to-image generation tasks using policy gradient methods (DDPO) with reward functions for aesthetic quality and prompt alignment. The candidate does not address denoising-stepwise optimization for mathematical reasoning in diffusion language models, which is the core innovation of wd1++.

3. Streaming diffusion policy: Fast policy synthesis with variable noise diffusion models

URL: View paper

Brief Assessment

Streaming Diffusion Policy[58] focuses on accelerating robotic policy synthesis through variable noise diffusion in action trajectories for visuomotor control, not on leveraging intermediate completions in iterative denoising for language model reasoning tasks.

4. Inpainting-Guided Policy Optimization for Diffusion Large Language Models

URL: View paper

Brief Assessment

Inpainting Policy Optimization[7] focuses on using inpainting to guide exploration during RL training by inserting partial ground-truth traces, not on leveraging intermediate completions from the denoising process itself for policy optimization as in wd1++.

5. FIND: Fine-tuning Initial Noise Distribution with Policy Optimization for Diffusion Models

URL: View paper

Brief Assessment

FIND Initial Noise[56] focuses on optimizing the initial noise distribution in diffusion models for image/video generation tasks, not on leveraging intermediate completions during iterative denoising for language model reasoning as in wd1++.

6. Diffusion policy policy optimization

URL: View paper

Brief Assessment

Diffusion Policy Optimization[53] focuses on fine-tuning diffusion-based policies for continuous control and robot learning tasks, not on denoising-stepwise policy optimization for reasoning in diffusion language models as proposed in the original paper's wd1++.

7. D2PPO: Diffusion Policy Policy Optimization with Dispersive Loss

URL: View paper

Brief Assessment

D2PPO Dispersive Loss[55] focuses on robotic manipulation using diffusion policies with dispersive loss regularization to prevent representation collapse, not on denoising-stepwise policy optimization for language model reasoning tasks as in the original paper.

---

8. Two-step diffusion policy deep reinforcement learning method for low-carbon multi-energy microgrid energy management

URL: View paper

Brief Assessment

Two-step Microgrid Management[59] focuses on energy management in multi-energy microgrids using a diffusion model-based policy with a two-step reward function, not on denoising-stepwise policy optimization for language model reasoning tasks.

---

9. Enhanced DACER Algorithm with High Diffusion Efficiency

URL: View paper

Brief Assessment

Enhanced DACER[57] focuses on improving diffusion efficiency in continuous control tasks through Q-gradient field guidance and temporal weighting mechanisms, not on leveraging intermediate completions from iterative denoising for language model policy optimization as in wd1++.

---

10. Step-Aware Policy Optimization for Reasoning in Diffusion Large Language Models

URL: View paper

Brief Assessment

Step-Aware Policy[9] focuses on process-based rewards and hierarchical reasoning structure in diffusion models, not on leveraging intermediate completions from the denoising process as in wd1++.

Appendix: Text Similarity Detection

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

References

- [0] wd1: Weighted Policy Optimization for Reasoning in Diffusion Language Models View paper
- [1] MMaDA: Multimodal Large Diffusion Language Models View paper
- [2] Mdpo: Overcoming the training-inference divide of masked diffusion language models View paper
- [3] Diffusion-based Large Language Models Survey View paper
- [4] Reinforcing the diffusion chain of lateral thought with diffusion language models View paper
- [5] Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review View paper
- [6] Consolidating Reinforcement Learning for Multimodal Discrete Diffusion Models View paper
- [7] Inpainting-Guided Policy Optimization for Diffusion Large Language Models View paper
- [8] LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models View paper
- [9] Step-Aware Policy Optimization for Reasoning in Diffusion Large Language Models View paper
- [10] Reinforcement Learning With LLMs Interaction for Distributed Diffusion Model Services View paper
- [11] Revolutionizing reinforcement learning framework for diffusion large language models View paper
- [12] Principled and Tractable RL for Reasoning with Diffusion Language Models View paper
- [13] Amortizing intractable inference in diffusion models for vision, language, and control View paper
- [14] Boundary-Guided Policy Optimization for Memory-efficient RL of Diffusion Large Language Models View paper
- [15] d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning View paper
- [16] SPG: Sandwiched Policy Gradient for Masked Diffusion Language Models View paper
- [17] VARD: Efficient and Dense Fine-Tuning for Diffusion Models with Value-based RL View paper
- [18] Training Diffusion Models with Reinforcement Learning View paper
- [19] DiffStitch: Boosting Offline Reinforcement Learning with Diffusion-based Trajectory Stitching View paper
- [20] User-centric interactive AI for distributed diffusion model-based AI-generated content View paper
- [21] Enhancing LLM QoS through Cloud-Edge Collaboration: A Diffusion-based Multi-Agent Reinforcement Learning Approach View paper
- [22] Enhancing reasoning for diffusion llms via distribution matching policy optimization View paper
- [23] Inference-Time Alignment Control for Diffusion Models with Reinforcement Learning Guidance View paper
- [24] Improving Reasoning for Diffusion Language Models via Group Diffusion Policy Optimization View paper
- [25] DiFFPO: Training Diffusion LLMs to Reason Fast and Furious via Reinforcement Learning View paper
- [26] Reasoning physical video generation with diffusion timestep tokens via reinforcement learning View paper
- [27] Learning Unmasking Policies for Diffusion Language Models View paper
- [28] Unifying Modern AI with Robotics: Survey on MDPs with Diffusion and Foundation Models View paper
- [29] Hybrid RAG-Empowered Multimodal LLM for Secure Data Management in Internet of Medical Things: A Diffusion-Based Contract Approach View paper
- [30] Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning View paper
- [31] MMaDA-Parallel: Multimodal Large Diffusion Language Models for Thinking-Aware Editing and Generation View paper
- [32] Diffusion-based Reinforcement Learning via Q-weighted Variational Policy Optimization View paper
- [33] LLM-Assisted Red Teaming of Diffusion Models through" Failures Are Fated, But Can Be Faded" View paper
- [34] Diffusion augmented agents: A framework for efficient exploration and transfer learning View paper
- [35] ReflectDiffu:Reflect between Emotion-intent Contagion and Mimicry for Empathetic Response Generation via a RL-Diffusion Framework View paper
- [36] Learning profitable NFT image diffusions via multiple visual-policy guided reinforcement learning View paper
- [37] d2: Improved Techniques for Training Reasoning Diffusion Language Models View paper
- [38] Lookahead Unmasking Elicits Accurate Decoding in Diffusion Language Models View paper
- [39] UniRL-Zero: Reinforcement Learning on Unified Models with Joint Language Model and Diffusion Model Experts View paper

- [40] Don't Settle Too Early: Self-Reflective Remasking for Diffusion Language Models View paper
- [41] LLM-I: LLMs are Naturally Interleaved Multimodal Creators View paper
- [42] Self-Reflective Reinforcement Learning for Diffusion-based Image Reasoning Generation View paper
- [43] ReconDreamer-RL: Enhancing Reinforcement Learning via Diffusion-based Scene Reconstruction View paper
- [44] d-TreeRPO: Towards More Reliable Policy Optimization for Diffusion Language Models View paper
- [45] Comparative Evaluation of Reasoning and Inference in LLM-Based and Diffusion-Based Approaches View paper
- [46] d1: Scaling Reasoning in Diï¬usion Large Language Models via Reinforcement Learning View paper
- [47] Dream-Coder 7B: An Open Diffusion Language Model for Code View paper
- [48] Make Imagination Clearer! Stable Diffusion-based Visual Imagination for Multimodal Machine Translation View paper
- [49] Modular Diffusion Policy Training: Decoupling and Recombining Guidance and Diffusion for Offline RL View paper
- [50] TARAD: Task-Aware Robot Affordance-Centric Diffusion Policy Learned From LLM-Generated Demonstrations View paper
- [51] Advantage weighted matching: Aligning rl with pretraining in diffusion models View paper
- [52] Diffusion Language Model with Query-Document Relevance for Query-Focused Summarization View paper
- [53] Diffusion policy policy optimization View paper
- [54] Efficient Online Reinforcement Learning for Diffusion Policy View paper
- [55] D2PPO: Diffusion Policy Policy Optimization with Dispersive Loss View paper
- [56] FIND: Fine-tuning Initial Noise Distribution with Policy Optimization for Diffusion Models View paper
- [57] Enhanced DACER Algorithm with High Diffusion Efficiency View paper
- [58] Streaming diffusion policy: Fast policy synthesis with variable noise diffusion models View paper
- [59] Two-step diffusion policy deep reinforcement learning method for low-carbon multi-energy microgrid energy management View paper