
Enhancing Reasoning for Diffusion LLMs via Distribution Matching Policy Optimization

Yuchen Zhu^{*1} Wei Guo^{*1} Jaemoo Choi¹ Petr Molodyk¹ Bo Yuan¹ Molei Tao^{†1} Yongxin Chen^{†1}

Abstract

Diffusion large language models (dLLMs) are promising alternatives to autoregressive large language models (AR-LLMs), as they potentially allow higher inference throughput. Reinforcement learning (RL) is crucial to enabling dLLMs to achieve performance comparable to that of AR-LLMs on important tasks, such as reasoning. However, RL algorithms well-suited to dLLMs' unique characteristics have yet to be developed. This paper proposes **Distribution Matching Policy Optimization (DMPO)**, a principled and theoretically grounded RL fine-tuning method specifically designed to enhance the reasoning capabilities of dLLMs by matching the dLLM policy distribution to the optimal, reward-tilted one through cross-entropy optimization. We identify a key implementation challenge with small training batch sizes and propose several effective solutions based on a novel weight baseline subtraction technique. DMPO exhibits superior performance on multiple reasoning benchmarks without supervised fine-tuning, achieving up to a 39.63 percentage-point improvement in accuracy over prior non-DMPO RL baselines and 67.97 percentage points over the base model, underscoring the effectiveness of the distribution-matching framework. Our code is available at <https://github.com/yuchen-zhu-zyc/DMPO>.

1. Introduction

Autoregressive large language models (AR-LLMs) have demonstrated remarkable capabilities in addressing sophisticated reasoning tasks, such as solving challenging math

^{*}Equal contribution [†]Equal advising ¹Georgia Institute of Technology. Correspondence to: Yuchen Zhu <yzhu738@gatech.edu>, Yongxin Chen <yongchen@gatech.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

questions and completing coding tasks (Jaech et al., 2024; Anthropic, 2025; Guo et al., 2025a; Novikov et al., 2025; Kimi Team et al., 2025). While these models form their amazing capabilities from pretraining on massive text corpora, the main powerhouse behind the success is scaling the post-training phase with reinforcement learning (RL) techniques, such as Proximal Policy Optimization (PPO, Schulman et al. (2017)) and Group Relative Policy Optimization (GRPO, Shao et al. (2024)), which enhance model abilities through exploration of reward functions and go beyond static datasets. While possessing extraordinary competence, AR-LLMs are known to be expensive for inference due to their sequential, fixed left-to-right generation order, which currently prohibits large-scale deployment.

To address such issues, diffusion large language models (dLLMs) have been investigated as an alternative to the AR models. Unlike AR-LLMs, dLLMs iteratively refine a sequence from masked inputs, enabling any-order generation, and have shown promising performance on text modeling tasks. dLLMs such as LLaDA (Nie et al., 2025b) and Dream (Ye et al., 2025), have demonstrated competitive performances on many tasks compared to similar-size AR baselines. Recently, commercial models such as Mercury (Inception Labs et al., 2025) and Gemini Diffusion (DeepMind) have demonstrated the ability to achieve significantly higher inference throughput without sacrificing generation quality, suggesting dLLMs as a promising direction for language modeling. However, one question that remains largely unanswered is how to transfer the success of RL on LLMs to dLLMs, thereby further scaling up their capability.

Designing RL algorithms for dLLMs faces two major challenges. Due to the bidirectional nature of dLLMs, estimating the log probability of generated sequences is more expensive than for AR models, making it less favorable to naively adapt LLM post-training algorithms like GRPO to dLLMs, which heavily rely on this estimate. The GRPO-style algorithms also do not leverage dLLM's unique characteristic of having a *forward* noising process, as they are backward-only algorithms when using generated rollouts. Moreover, existing RL frameworks for enhancing LLM reasoning capabilities overly focus on reward maximization (Guo et al., 2025a; Liu et al., 2025b; Zheng et al., 2025a). By targeting

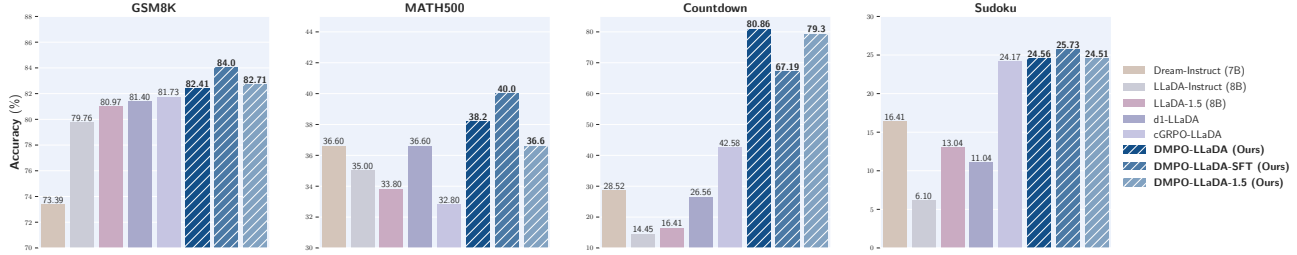


Figure 1. Performance on reasoning benchmarks evaluated with generation length 256. DMPO consistently achieves the best performance across dLLMs, outperforming prior RL baselines.

only the reward mode, these approaches do not fully leverage dLLMs’ potential to generate diverse responses, given the task’s random-order nature (Gong et al., 2026).

To jointly address these challenges, we propose **Distribution Matching Policy Optimization (DMPO)**, a principled and efficient RL fine-tuning method specifically designed for dLLMs. DMPO is designed on a novel framework grounded in stochastic optimal control (SOC), which shifts away from the conventional reward-maximization paradigm and targets the goal of matching the entire reward-tilted policy distribution. This enables the model to explore diverse, high-quality reasoning paths and responses during training, addressing concerns about over-focusing on absolute reward values and modes. In addition, DMPO training leverages importance sampling and a novel weighted denoising cross-entropy (WDCE) loss, which has the key advantage of operating in an **off-policy** manner, enabling the use of replay buffers to improve sample efficiency. More importantly, WDCE is a **forward-only** objective that relies solely on the obtained clean samples and the inexpensive, forward-noising process unique to dLLMs. DMPO largely decouples from rollout trajectories, potentially enabling it to achieve greater speed-ups than other dLLM RL algorithms when employed with fast inference techniques.

Contributions (I) We propose a novel RL learning framework for dLLMs that targets distribution matching rather than reward maximization (Sec. 3.1). (II) We propose Distribution Matching Policy Optimization (DMPO), a principled, theoretically-grounded fine-tuning strategy for enhancing dLLM’s reasoning capabilities, supported by importance sampling and weighted denoising cross-entropy (Sec. 3.2). (III) We identify a special challenge that occurred for WDCE due to the use of a limited training batch size, and propose two novel techniques to address it: weight baseline subtraction (Sec. 3.3) and weighted direct discriminative optimization (Sec. 3.4). (IV) DMPO exhibits superior performances on multiple reasoning benchmarks without supervised fine-tuning (SFT), with an accuracy improvement up to 39.63 percentage points over prior non-DMPO RL baselines and 67.97 percentage points over the base model, being top-performing across bi-directional dLLMs (Sec. 4).

2. Preliminaries

2.1. Masked Diffusion Models for Language Modeling

The **masked (discrete) diffusion models (MDM)** (Lou et al., 2024; Ou et al., 2025; Sahoo et al., 2024; Shi et al., 2024; Zheng et al., 2025f) is a novel method for learning high-dimensional categorical distributions with application to text (Nie et al., 2025b), images (Chang et al., 2022; Bai et al., 2025), DNAs (Hayes et al., 2025), etc. Essentially, it learns the one-dimensional conditional distributions of the data given any subset of observed dimensions. Suppose the data are finite-length sequences with vocabulary $\mathcal{V} = \{1, 2, \dots, V\}$. Include the mask token M into the \mathcal{V} and let $\bar{\mathcal{V}} = \{1, 2, \dots, V, M\}$. The MDM takes a partially masked sequence $\mathbf{x} = (x_1, \dots, x_D) \in \bar{\mathcal{V}}^D$ as an input, and outputs $\pi_\theta(\mathbf{x}) \in \mathbb{R}^{D \times V}$, whose (d, u) -th entry $\pi_\theta(\mathbf{x})_{d,u}$ is set to $1_{x_d=u}$ if $x_d \neq M$, and if $x_d = M$, is trained to approximate the conditional probability

$$\Pr_{\mathbf{x} \sim p_{\text{data}}}(X_d = u | \mathbf{X}_{UM} = \mathbf{x}_{UM}),$$

where $\mathbf{x}_{UM} = (x_d : x_d \neq M)$. By definition, we assume each row of $\pi_\theta(\mathbf{x})$ is a valid probability vector. The probability of an unmasked sequence $\mathbf{x} \in \mathcal{V}^D$ under the MDM π_θ is defined through **random-order autoregressive (ROAR) generation**: choosing a uniformly random order of the D positions, and autoregressively sampling each position conditional on the previously sampled ones. Formally, let σ be a uniformly random permutation of $\{1, \dots, D\}$, then

$$p_\theta(\mathbf{x}) = \mathbb{E}_\sigma p_\theta(\mathbf{x}; \sigma), p_\theta(\mathbf{x}; \sigma) = \prod_{d=1}^D \pi_\theta(x_{\sigma_d} | \mathbf{x}_{\sigma_{<d}}). \quad (1)$$

Here, $\pi_\theta(x_{\sigma_d} | \mathbf{x}_{\sigma_{<d}})$ means input \mathbf{x} with all positions except $\sigma_{<d} = \{\sigma_1, \dots, \sigma_{d-1}\}$ masked into the MDM and take the output at position (σ_d, x_{σ_d}) .

The standard way to train an MDM given i.i.d. samples from p_{data} is to minimize the **denoising cross-entropy (DCE)** loss $\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathcal{L}_\theta(\mathbf{x})$, which involves the following definition of the (negative) **evidence lower bound (ELBO)** \mathcal{L}_θ :

$$\begin{aligned}
 -\log p_\theta(\mathbf{x}) &= -\log \mathbb{E}_\sigma p_\theta(\mathbf{x}; \sigma) \leq -\mathbb{E}_\sigma \log p_\theta(\mathbf{x}; \sigma) \\
 &= \mathbb{E}_{m \sim \text{Unif}\{1, \dots, |\mathbf{x}|\}} \left[\frac{|\mathbf{x}|}{m} \mathbb{E}_{\mu_m(\tilde{\mathbf{x}}|\mathbf{x})} \sum_{d: \tilde{x}_d=M} -\log \pi_\theta(\tilde{\mathbf{x}})_{d, x_d} \right] \\
 &=: \mathcal{L}_\theta(\mathbf{x}), \tag{2}
 \end{aligned}$$

where the transition distribution $\mu_m(\cdot|\mathbf{x})$ means to sample a uniformly random subset of $\{1, \dots, |\mathbf{x}|\}$ of size m and mask the corresponding entries in \mathbf{x} , and $|\mathbf{x}|$ is the length of \mathbf{x} . The proof of the last equation can be found in Uria et al. (2016); Ou et al. (2025).

When applying to text data, the MDM is also referred to as the **diffusion large language model (dLLM)** (Nie et al., 2025b; Ye et al., 2025; Inception Labs et al., 2025; Song et al., 2025). For the purpose of reasoning, we typically write $\mathbf{x} = (\mathbf{q}, \mathbf{o})$, where \mathbf{q} is the **prompt** (or query, which is always assumed to contain no mask state) and \mathbf{o} is the **response** (or output). We use $\pi_\theta(\mathbf{o}|\mathbf{q}) \in \mathbb{R}^{|\mathbf{o}| \times V}$ to denote the policy model output of the dLLM given a prompt \mathbf{q} and a partially masked response \mathbf{o} . The conditional sequence probability of a clean model \mathbf{o} given a prompt \mathbf{q} , denoted as $p_\theta(\mathbf{o}|\mathbf{q})$, is similarly defined through (1), where we now use notations $p_\theta(\mathbf{o}|\mathbf{q}; \sigma)$ and $\pi_\theta(o_d|\mathbf{q}, \mathbf{o}_{<d})$ to emphasize the dependence on the prompt \mathbf{q} . The negative ELBO will be written as $\mathcal{L}_\theta(\mathbf{o}|\mathbf{q})$.

2.2. Reinforcement Learning for Enhancing Reasoning

We first present the **Group Relative Policy Optimization (GRPO)**, Shao et al. (2024) method for LLMs, which is the basis of most of the existing RL methods for dLLMs. Given a pretrained LLM with policy π_{ref} that samples from the distribution $p_{\text{ref}}(\mathbf{o}|\mathbf{q}) = \prod_{d=1}^{|\mathbf{o}|} \pi_{\text{ref}}(o_d|\mathbf{q}, \mathbf{o}_{<d})$, a reward function $r : (\mathbf{q}, \mathbf{o}) \mapsto \mathbb{R}$, a set of prompts \mathcal{D} , and a regularization parameter $\alpha \geq 0$, each step of the GRPO aims to solve the following problem: sample prompt $\mathbf{q} \sim \mathcal{D}$ and rollouts $\mathbf{o}^{(1:G)} \stackrel{\text{i.i.d.}}{\sim} p_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})$, then, maximize

$$\mathbb{E} \left\{ \frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}^{(i)}|} \sum_{d=1}^{|\mathbf{o}^{(i)}|} \left[\min \left(\rho_d^{(i)} A_i, \text{clip}(\rho_d^{(i)})_{1 \pm \epsilon} A_i \right) - \alpha \text{KL}(p_\theta(\mathbf{o}^{(i)}|\mathbf{q}) \| p_{\text{ref}}(\mathbf{o}^{(i)}|\mathbf{q})) \right] \right\}, \tag{3}$$

where the advantages¹ are $A_i = r(\mathbf{q}, \mathbf{o}^{(i)}) - \text{mean}(r(\mathbf{q}, \mathbf{o}^{(1:G)}))$, the per-token probability ratios are $\rho_d^{(i)} = \frac{\pi_\theta(o_d^{(i)}|\mathbf{q}, \mathbf{o}_{<d}^{(i)})}{\pi_{\theta_{\text{old}}}(o_d^{(i)}|\mathbf{q}, \mathbf{o}_{<d}^{(i)})}$, and the KL regularization term is estimated similarly by the per-token probability ratios between π_θ and π_{ref} . The clipping threshold ϵ prevents overly large policy updates.

¹As suggested by Liu et al. (2025b), we list here the version without normalization by standard deviation.

While (3) works well for LLMs, it is not directly applicable to dLLMs due to mismatch between the *dLLM policy (model output)* $\pi_\theta(\mathbf{o}|\mathbf{q})$ and the *sequence likelihood* $p_\theta(\mathbf{o}|\mathbf{q})$: unlike in LLMs where these two quantities are easily connected through the chain rule, it is generally non-trivial to compute the per-token probability given the dLLM model output, and only ELBO (2) is available as a surrogate. To tackle this issue, diffu-GRPO (Zhao et al., 2025a) proposed to fully mask all response positions and partially masks the prompt \mathbf{q} , and feed this sequence into the model to obtain the approximate probability $p_\theta(o_d|\mathbf{q})$. Next, the sequence probability $p_\theta(\mathbf{o}|\mathbf{q})$ is approximated by mean-field decomposition: $p_\theta(\mathbf{o}|\mathbf{q}) \approx \prod_{d=1}^{|\mathbf{o}|} p_\theta(o_d|\mathbf{q})$. Such approximations do not capture correlations among different positions in the response, leading to imprecision. A similar technique is employed in coupled-GRPO (cGRPO) for code generation tasks in Gong et al. (2026).

3. Distribution Matching Policy Optimization

3.1. Reward Maximization \rightarrow Distribution Matching

To incentivize the reasoning capabilities of large language models, reward-maximizing reinforcement learning finetuning algorithms, such as TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017), and GRPO (Shao et al., 2024), are often employed, with an additional reward entropy regularization term that penalizes the deviation of the model from the pretrained one. This process amounts to solving the following optimization problem,

$$\max_{\theta} \mathbb{E}_{\mathbf{q} \sim \mathcal{D}} \left[\mathbb{E}_{p_\theta(\mathbf{o}|\mathbf{q})} r(\mathbf{q}, \mathbf{o}) - \alpha \text{KL}(p_\theta(\cdot|\mathbf{q}) \| p_{\text{ref}}(\cdot|\mathbf{q})) \right]. \tag{4}$$

However, existing techniques over-focus on finding and optimizing the **reward mode** and adopt many heuristic techniques to accelerate the mode searching process, neglecting the exploration of the entire distribution landscape, and often result in model mode collapse or reward hacking, causing the model to produce undesirable responses (Weng, 2024). A simple fix to this issue and to encourage diverse model responses is to enforce the optimality of the target policy distribution during the training. It can be shown that the optimal sequence distribution that solves the problem (4) is the following **reward-tilted distribution**:

$$p_*(\mathbf{o}|\mathbf{q}) = \frac{1}{Z(\mathbf{q})} p_{\text{ref}}(\mathbf{o}|\mathbf{q}) e^{r(\mathbf{q}, \mathbf{o})/\alpha}, \tag{5}$$

where $Z(\mathbf{q}) = \sum_{\mathbf{o}} p_{\text{ref}}(\mathbf{o}|\mathbf{q}) e^{r(\mathbf{q}, \mathbf{o})/\alpha}$. That is to say, we

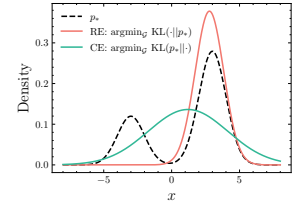


Figure 2. Illustration of relative entropy (mode-seeking) and cross-entropy (mass-covering) for fitting a target p_* (\mathcal{G} is the set of Gaussian distributions).

want to use the optimal sequential distribution $p_*(\mathbf{o}|\mathbf{q})$ as the **supervision signal** throughout the learning process, so that we can learn a dLLM policy π_θ which produces a sequence distribution p_θ matching p_* . We can thus obtain a policy that not only explores the dominant reward mode but is also guaranteed to sample other high-reward trajectories with a likelihood proportional to their reward values. This motivates us to consider the following task:

Policy Distribution Matching Learning: Given a pretrained dLLM policy $\pi_{\text{ref}}(\mathbf{o}|\mathbf{q})$ that samples from a distribution $p_{\text{ref}}(\mathbf{o}|\mathbf{q})$, a reward function $r : (\mathbf{q}, \mathbf{o}) \mapsto \mathbb{R}$, a set of prompts \mathcal{D} , and temperature $\alpha > 0$, learn a dLLM policy $\pi_\theta(\mathbf{o}|\mathbf{q})$ to produce the desired optimal distribution $p_*(\mathbf{o}|\mathbf{q})$ (5) by optimizing the following objective:

$$\min_{\pi_\theta} \mathbb{E}_{\mathbf{q} \sim \mathcal{D}} \mathcal{F}(p_\theta(\cdot|\mathbf{q}), p_*(\cdot|\mathbf{q})). \quad (6)$$

Here, \mathcal{F} is a class of functionals such that $\text{argmin}_p \mathcal{F}(p, p_*) = p_*$. Note that the original entropy-regularized entropy optimization problem is equivalent to choosing \mathcal{F} to be the reverse KL between p and p_* , i.e., $\mathcal{F}(p_\theta, p_*) = \text{KL}(p_\theta \| p_*) = \mathbb{E}_{p_\theta} [\log \frac{p_\theta}{p_*}]$. While this objective, in theory, can also yield the same optimal distribution with the desired property, reverse KL is widely known to be *mode-seeking*, i.e., it tends to match the dominant mode in p_* while potentially neglecting other modes, which may lead to reward hacking.

To address this issue, we consider a series of new objectives \mathcal{F} with more desirable convergence guarantees that steadily lead to optimization towards the desired sequence distribution, and propose **Distribution Matching Policy Optimization (DMPO)** (Alg. 1), which targets matching the entire reward-tilted policy distribution. In Sec. 3.2, we introduce **weighted denoising cross-entropy (WDCE)**, a **scalable** implementation of the forward KL using importance sampling. In Secs. 3.3 and 3.4, we discuss an important failure case of forward KL with **small training batch size**, and propose a series of novel techniques such as **weight baseline subtraction** (Sec. 3.3) and **weighted direct discriminative optimization** (Sec. 3.4) to address it.

3.2. Weighted Denoising Cross-entropy

Unlike the reverse KL objective considered by many existing works, which are known to be prone to mode seeking and collapse, one alternative choice is to use the forward KL divergence (or **cross-entropy, CE**) for the functional, i.e., $\mathcal{F}(p_\theta, p_*) = \text{KL}(p_* \| p_\theta)$, which tends to cover all the modes of the optimal distribution and can retain the response diversity. The CE loss is widely used in another domain, stochastic optimal control (SOC) (Domingo-Enrich et al., 2024; 2025), which is closely connected to our work. This

amounts to solving the following task,

$$\min_{\theta} \mathbb{E}_{\mathbf{q} \sim \mathcal{D}} \mathbb{E}_{p_*(\mathbf{o}|\mathbf{q})} \left[\log \frac{p_*(\mathbf{o}|\mathbf{q})}{p_\theta(\mathbf{o}|\mathbf{q})} \right]. \quad (7)$$

However, objective (7) is not directly amenable to practical implementation, as we do not have access to real samples from the p_* , nor can we exactly compute $\log p_*$ due to the presence of the unknown partition function $Z(\mathbf{q})$. To bypass this issue, we draw inspiration from the recent work masked diffusion neural sampler (MDNS, Zhu et al. (2025e)), which proposed a training framework for learning a neural sampler based on MDM with stochastic optimal control and cross-entropy minimization. While targeting a different task, the core of MDNS resides in solving the same distribution matching problem with cross-entropy loss, and it proposed a practically implementable and scalable variant of (7) named **weighted denoising cross-entropy (WDCE)** loss. The central idea is to introduce a reference policy and leverage *importance sampling* to treat i.i.d. samples as importance-weighted samples from p_* . Taking advantage of this approach, we now derive WDCE for dLLM policy learning.

First, given the relationship between the policy output and sequence distribution of the masked dLLM (1), it is clear that we can match the correct target sequence distribution $p_*(\mathbf{o}|\mathbf{q})$ as long as we train $p_\theta(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})$ to match the *order-specific* ones, i.e., $p_*(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})$, given by

$$p_*(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma}) = \frac{1}{Z(\mathbf{q})} p_{\text{ref}}(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma}) e^{r(\mathbf{q}, \mathbf{o})/\alpha}. \quad (8)$$

Leveraging this fact, given any prompt \mathbf{q} , we can express the cross-entropy loss as follows:

$$\begin{aligned} \text{KL}(p_*(\cdot|\mathbf{q}) \| p_\theta(\cdot|\mathbf{q})) &= \mathbb{E}_{p_*(\mathbf{o}|\mathbf{q})} [-\log p_\theta(\mathbf{o}|\mathbf{q})] + \text{const} \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{p_*(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})} [-\log p_\theta(\mathbf{o}|\mathbf{q})] + \text{const} \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})} \frac{p_*(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})}{p_v(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})} [-\log p_\theta(\mathbf{o}|\mathbf{q})] + \text{const}, \end{aligned} \quad (9)$$

where p_v is the sequence probability under a reference policy model v that does not involve gradient computation, and in practice, one often chooses $v \leftarrow \bar{\theta} := \text{stopgrad}(\theta)$ to be a copy of the policy model detached from the computation graph, and periodically synchronizes with the current model policy p_θ , which is also commonly referred to as $p_{\theta_{\text{old}}}$ in the literature. The importance weight $w(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma}) := \frac{p_*(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})}{p_v(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})}$ captures the mismatch between p_v and p_* and ensures the mathematical correctness of the objective, and $\log p_\theta(\mathbf{o}|\mathbf{q})$ is an intractable sequence log probability under the current dLLM policy. We discuss the computation of these two components in parallel below.

Importance weight $w(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})$ We simplify it with the pretrained model and the reward:

$$w(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma}) = \frac{1}{Z(\mathbf{q})} \frac{p_{\text{ref}}(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})}{p_v(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})} e^{\frac{r(\mathbf{q},\mathbf{o})}{\alpha}}$$

$$\propto \exp\left(\frac{r(\mathbf{q},\mathbf{o})}{\alpha} + \log \frac{p_{\text{ref}}(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})}{p_v(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})}\right) =: e^{\ell(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})}. \quad (10)$$

Recall that the order-specific probability of a sequence is computed via (1). To ensure that the sample distribution after importance sampling is valid and normalized, we keep track of the **log weights** $\ell(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})$, and take softmax among those corresponding to the same prompt \mathbf{q} to compute the real weight $w(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})$. This is equivalent to estimating the unknown partition function $Z(\mathbf{q})$ using an empirical estimator of the following expectation:

$$Z(\mathbf{q}) = \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})} \left[\frac{p_{\text{ref}}(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})}{p_v(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})} e^{r(\mathbf{q},\mathbf{o})/\alpha} \right].$$

The need to estimate partition functions is common in RL algorithms for LLM, such as in GflowNet (Bengio et al., 2021; Kimi Team et al., 2025). In contrast to these approaches that learn such functions independently, our estimation approach is training-free and more efficient.

Sequence log probability $\log p_v(\mathbf{o}|\mathbf{q})$ Unlike the case of LLM, the exact sequence log probability is intractable in dLLM due to the presence of expectation over the random order $\boldsymbol{\sigma}$. However, similar to the training of dLLM, we can leverage the negative ELBO (2) as a surrogate. Combined with the importance weight $w(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})$, we introduce the **weighted denoising cross-entropy (WDCE)** loss for dLLM policy distribution matching:

$$\min_{\theta} \mathbb{E}_{\mathbf{q} \sim \mathcal{D}} \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma})} \left\{ w(\mathbf{o}|\mathbf{q};\boldsymbol{\sigma}) \mathbb{E}_{m \sim \text{Unif}\{1, \dots, |\mathbf{o}|\}} \left[\frac{|\mathbf{o}|}{m} \mathbb{E}_{\mu_m(\tilde{\mathbf{o}}|\mathbf{o})} \sum_{d: \tilde{\mathbf{o}}_d = \mathbf{M}} -\log \pi_{\theta}(\tilde{\mathbf{o}})_{d, \mathbf{o}_d} \right] \right\}. \quad (11)$$

Notably, this loss highly resembles the DCE loss used in pre-training and the supervised fine-tuning (SFT) phase of dLLM. One major difference is that instead of using i.i.d. samples from p_* , we use *importance sampling* to weight samples from p_v and obtain a valid training objective with theoretical guarantees. WDCE differs significantly from other popular RL training techniques, such as PPO/GRPO, in the following two key aspects:

WDCE is an off-policy loss The WDCE loss remains valid as the model parameter θ gets updated, since both the sampling policy p_v and the important sampling target policy p_* are independent of the current model policy p_{θ} . This allows us to save generated rollouts in a replay buffer and reuse them across multiple training updates, without worrying excessively about numerical instability, thereby improving sample efficiency. On the other hand, for on-policy methods, to use a replay buffer, one would need

to estimate importance weights with respect to the current model policy $p_{\theta}(\mathbf{o}|\mathbf{q})$, i.e., $\frac{p_{\theta}(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})}$. Different from the case of LLM where such estimation can be done in one model forward pass, an accurate estimation in dLLM **per training update** is expensive, rendering the on-policy method less efficient. Moreover, for large models, where rollout generation and sequence likelihood estimation are typically handled by different implementations (such as vLLM and FSDP), this could lead to more nuanced, hard-to-detect biases that secretly undermine the algorithm’s performance (Yao et al., 2025). With WDCE, we are largely free of such concerns.

WDCE is a forward loss Unlike GRPO-style algorithms that typically require tracking the entire rollout trajectories, WDCE leverages the forward (noising) process during training, which is a characteristic *unique* to **diffusion** LLMs. Once we obtain the final samples and their associated weights, we can discard the trajectories and perform training using the cheap forward process by randomly masking the data. This implies that the training speed with WDCE largely depends on the model’s inference speed. With the advances of dLLM efficient inference techniques such as fast decoding algorithms and KV-cache techniques (Ma et al., 2025; Hu et al., 2025; Wu et al., 2026; Liu et al., 2025c), WDCE could also enjoy a great boost in efficiency. This method also effectively leverages dLLM’s potential to surpass LLMs in inference throughput, distinguishing it from other RL baselines that merely adapt LLM algorithms to dLLM. We defer a more detailed discussion of such properties to App. B.2.

Stochastic optimal control (SOC) perspective While we developed WDCE through the lens of distribution matching (6), it can also be derived from a path-measure SOC viewpoint, following Wang et al. (2025a); Zhu et al. (2025e); we sketch the argument here and defer details to App. B.1. The random-order autoregressive sampling of π_{θ} can be viewed as a continuous-time Markov chain (CTMC) on $\bar{\mathcal{V}}^D$ with rate matrix $Q_t^{\theta}(\mathbf{x}, \mathbf{x}^{d \leftarrow n}) = \gamma(t) \pi_{\theta}(\mathbf{x})_{d, n} \mathbf{1}_{x^d = \mathbf{M}}$, inducing a **path measure** \mathbb{P}^{θ} whose terminal distribution coincides with p_{θ} . To produce p_* at the terminal time, we define the **target path measure**

$$\mathbb{P}^*(\boldsymbol{\xi}) := \frac{1}{Z} \mathbb{P}^{\text{ref}}(\boldsymbol{\xi}) e^{r(\boldsymbol{\xi}_1)/\alpha},$$

which preserves the pretrained transition dynamics while tilting its terminal distribution to exactly p_* in (5). A standard CTMC Radon–Nikodým calculation (see Lem. B.1) shows that diagonal and integral contributions cancel, leaving the closed-form path-level log-weight $\log \frac{d\mathbb{P}^*}{d\mathbb{P}^{\theta}}(\boldsymbol{\xi}) = W^{\theta}(\boldsymbol{\xi}) - \log Z$, where $W^{\theta}(\boldsymbol{\xi})$ is the trajectory log-weight defined in (22) and depends only on the terminal reward $r(\boldsymbol{\xi}_1)$ and on along-trajectory policy ratios $\frac{\pi_{\text{ref}}}{\pi_{\theta}}$ at the jump

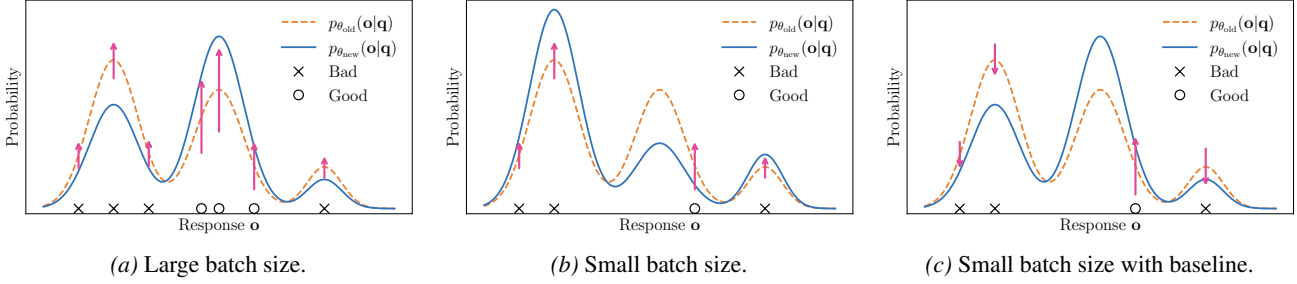


Figure 3. Demonstration of the effect of weight baseline. The orange and blue curves represent the probability $p_{\theta}(\mathbf{o}|\mathbf{q})$ before and after update, and the magenta arrows represent the weights. (a) When the batch size is large, the distribution mode coverage is good. Though bad responses have positive weights, the correct ones have larger weights to push the distribution updates in the right direction. (b) When the batch size is small, some modes (e.g., the good one in the middle) may not be sampled. Without **weight baseline subtraction**, the dominant positive weights of the bad responses lead to wrong update directions. (c) With **weight baseline subtraction**, the bad responses will appropriately be penalized, leading to the desired update direction.

times. Minimizing $\text{KL}(\mathbb{P}^* \parallel \mathbb{P}^{\theta})$ by importance sampling from a stop-gradient copy \mathbb{P}^v gives

$$\text{KL}(\mathbb{P}^* \parallel \mathbb{P}^{\theta}) = \mathbb{E}_{\mathbb{P}^v(\xi)} \frac{1}{Z} e^{W^v(\xi)} \mathcal{L}_{\theta}(\xi_1) + \text{const},$$

which coincides *exactly* with the WDCE loss (11). Crucially, \mathcal{L}_{θ} is the *exact* negative log-density of \mathbb{P}^{θ} on the path space, so the ELBO approximation used at the sequence level in (9) corresponds to *precise* KL matching at the path level, justifying the validity of such surrogate.

3.3. Effective Training with Negative Gradient Insertion

While theoretically, minimizing the WDCE loss (11) provably leads to convergence of the model sequence distribution to $p_*(\mathbf{o}|\mathbf{q})$, this could face practical issues due to the often limited number of rollouts generated per prompt. Ideally, we would want to increase the likelihood of “good” responses while decreasing the likelihood of “bad” responses. However, with WDCE, any response \mathbf{o} will be associated with a *positive* weight $w(\mathbf{o}|\mathbf{q}; \sigma)$ due to the softmax operation, which may lead to ineffective learning in the low-batch-size scenario.

We note that this issue does not arise when the batch size is sufficiently large for the following reason. When having a large batch of diverse responses that make up a good coverage of the sample space, despite having all positive weights, since the model cannot increase likelihood on all responses (as the probabilities sum up to 1), the “bad” responses will be automatically and implicitly penalized due to not having larger weights than the “good” responses.

When the batch size is small, the scenario is different as is illustrated in Fig. 3. In such a case, the model will tend to **promote both “good” and “bad” responses** due to the positive weights, and potentially penalize the likelihood of other unseen responses to maintain a valid probability distribution. This could be detrimental to achieving distribution

matching, as these unseen responses may have high reward values and correspond to an undiscovered distribution mode.

To address this issue, we inject negative gradient (Ren & Sutherland, 2025; Deng et al., 2025) by designing a **weight baseline** and subtract it from the obtained weights to facilitate an effective reinforcement on the good samples, i.e.,

$$w_{\text{real}}(\mathbf{o}|\mathbf{q}; \sigma) = w(\mathbf{o}|\mathbf{q}; \sigma) - w_{\text{base}}(\mathbf{o}|\mathbf{q}; \sigma). \quad (12)$$

This approach resembles that adopted by PPO/GRPO. However, unlike these methods, we rate responses based on the log weights $\ell(\mathbf{o}|\mathbf{q}; \sigma)$, where larger values indicate better alignment with the target optimal distribution. As a consequence, we promote responses that are more likely to be sampled from p_* and penalize those that are less likely. Based on this perspective, we consider the following three methods for choosing $w_{\text{base}}(\mathbf{o}|\mathbf{q}; \sigma)$.

Group weight baseline When the dLLM policy is close to optimal, the original log weight $\ell(\mathbf{o}|\mathbf{q}; \sigma)$ should behave approximately like constants for a group of different responses $\{\mathbf{o}^{(n)}\}_{1 \leq n \leq N}$, leading to nearly uniform weight value for $\{w(\mathbf{o}^{(n)}|\mathbf{q}; \sigma^{(n)})\}_{1 \leq n \leq N}$ after group softmax. We can thus choose the baseline as 1 to encourage convergence to this optimal situation:

$$w_{\text{base}}(\mathbf{o}^{(n)}|\mathbf{q}; \sigma^{(n)}) = 1. \quad (13)$$

Individual weight baseline We can also consider the individual weight value of each response. For samples with smaller log weights, a stronger penalization is more desirable. A natural, adaptive way to design penalization strength is to use softmax over the log weights with *negative reward*: define $\ell_{-}(\mathbf{o}|\mathbf{q}; \sigma) := -\frac{r(\mathbf{q}, \mathbf{o})}{\alpha} + \log \frac{p_{\text{ref}}(\mathbf{o}|\mathbf{q}; \sigma)}{p_v(\mathbf{o}|\mathbf{q}; \sigma)}$, and

$$w_{\text{base}}(\mathbf{o}^{(n)}|\mathbf{q}; \sigma^{(n)}) = \frac{N \exp(\ell_{-}(\mathbf{o}^{(n)}|\mathbf{q}; \sigma^{(n)}))}{\sum_k \exp(\ell_{-}(\mathbf{o}^{(k)}|\mathbf{q}; \sigma^{(k)}))}. \quad (14)$$

Note that this $w_{\text{base}}(\mathbf{o}|\mathbf{q}; \sigma)$ now corresponds to a *bad target distribution* given by $p_{*-}(\mathbf{o}|\mathbf{q}) \propto_o p_{\text{ref}}(\mathbf{o}|\mathbf{q}) e^{-r(\mathbf{q}, \mathbf{o})/\alpha}$,

which is tilted by the negative reward. The minus sign in the loss before $w_{\text{base}}(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})$ means we want to drive the dLLM policy away from this bad distribution.

Model weight baseline Finally, we can determine whether to promote or penalize specific responses by comparing $w(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})$ with the importance weight under the current model policy $p_\theta(\mathbf{o}|\mathbf{q})$, which pushes the model further towards the optimal $p_*(\mathbf{o}|\mathbf{q})$. Note that this does not incur additional computation overhead as we can estimate $\log p_{\bar{\theta}}(\mathbf{o}|\mathbf{q})$ using negative ELBO (2), which is already computed in the WDCE loss. Define $\ell_\theta(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma}) := \log \frac{p_{\bar{\theta}}(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})}{p_v(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})}$, and let

$$w_{\text{base}}(\mathbf{o}^{(n)}|\mathbf{q}; \boldsymbol{\sigma}^{(n)}) = \frac{N \exp(\ell_\theta(\mathbf{o}^{(n)}|\mathbf{q}; \boldsymbol{\sigma}^{(n)}))}{\sum_k \exp(\ell_\theta(\mathbf{o}^{(k)}|\mathbf{q}; \boldsymbol{\sigma}^{(k)}))}. \quad (15)$$

We remark that the group weight and model weight baselines (13) and (15) can also be interpreted as an *approximate variance reduction*. See App. B.3 for discussion.

3.4. Weighted Direct Discriminative Optimization

To explore the full potential of the distribution matching framework in (6), we also investigate other choices for the potential \mathcal{F} beyond the cross-entropy. One particularly interesting objective is the following **direct discriminative optimization (DDO)** loss:

$$\begin{aligned} \mathcal{F}(p_\theta(\cdot|\mathbf{q}), p_*(\cdot|\mathbf{q})) &= -\mathbb{E}_{p_*(\mathbf{o}|\mathbf{q})} \log \sigma \left(\log \frac{p_\theta(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})} \right) \\ &\quad - \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q})} \log \sigma \left(-\log \frac{p_\theta(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})} \right), \quad (16) \end{aligned}$$

where $\sigma(t) = 1/(1 + e^{-t})$. The global minimizer of (16) is also $p_*(\cdot|\mathbf{q})$, thus being a valid functional for distribution matching. For a more detailed justification, see App. B.4.

This is inspired by Zheng et al. (2025e), which proposed a GAN-like (Goodfellow et al., 2020) loss for SFT of vision models. One interesting trait of this objective is its natural incorporation of negative gradients for bad samples due to the GAN nature, as is shown in the analysis therein:

$$\begin{aligned} \nabla_\theta \mathcal{F}(p_\theta(\cdot|\mathbf{q}), p_*(\cdot|\mathbf{q})) &= \sum_{\mathbf{o}} \sigma \left(-\log \frac{p_\theta(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})} \right) \\ &\quad \cdot \left[(p_\theta(\mathbf{o}|\mathbf{q}) - p_*(\mathbf{o}|\mathbf{q})) \nabla_\theta \log p_\theta(\mathbf{o}|\mathbf{q}) \right]. \end{aligned}$$

From the expression, as the first term is always non-negative, and the boxed term applies a penalty for bad response \mathbf{o} , thus providing a gradient direction for increasing $p_\theta(\mathbf{o}|\mathbf{q})$. Leveraging this property, we adapt it for RL finetuning of dLLM and introduce **weighted direct discriminative optimization (WDDO)** loss, again dealing with p_* through *importance sampling*:

$$\begin{aligned} \mathcal{F}(p_\theta(\cdot|\mathbf{q}), p_*(\cdot|\mathbf{q})) &= -\mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})} \left[w(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma}) \right. \\ &\quad \left. \log \sigma \left(\log \frac{p_\theta(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})} \right) + \log \sigma \left(-\log \frac{p_\theta(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})} \right) \right], \end{aligned}$$

where $w(\mathbf{o}|\mathbf{q}; \boldsymbol{\sigma})$ is the importance weight defined in (10).

4. Experiments

Model and baselines We apply DMPO to LLaDA-8B-Instruct (Nie et al., 2025b), a SOTA fully-bidirectional dLLM not post-trained with RL techniques. To clearly demonstrate the potential of DMPO, we follow an R1-Zero-like training recipe (Guo et al., 2025a; Liu et al., 2025b) and apply DMPO directly to the LLaDA and Dream models without first performing SFT on curated datasets. We refer to the model obtained from this pipeline as **DMPO-LLaDA** and **DMPO-Dream**. We benchmark our method against a series of dLLM base models of comparable size, such as Dream-Instruct (7B, Ye et al. (2025)), LLaDA-Instruct (8B, Nie et al. (2025b)), and LLaDA-1.5 (8B, Zhu et al. (2025a)). We compare with d1 (Zhao et al., 2025a), a dLLM RL finetuning approach that combines both SFT and diffu-GRPO (an adapted version of GRPO), as well as cGRPO (Gong et al., 2026), wd1 (Tang et al., 2025c), and GDPO (Rojas et al., 2025a).

Experimental setups We explore four reasoning benchmarks: GSM8k, MATH500, Sudoku, and Countdown. For all pretrained dLLMs, we evaluate the latest available checkpoint. For d1 and cGRPO, we reproduce their results exactly following the provided guidelines. To ensure a fair comparison, we train DMPO-LLaDA and DMPO-Dream on the same datasets as d1 for each task with rollouts generated using a fixed sequence length of 256. Evaluations are conducted with zero-shot prompting using generation lengths 128, 256, and 512 for LLaDA-series models, and 256 and 512 for Dream.

DMPO incentivizes superior reasoning capabilities We report in Tabs. 1 and 2 the performance of DMPO together with that of base dLLMs and models obtained by existing RL post-training strategies. DMPO consistently outperforms both the base models and the models tuned with d1/cGRPO/wd1/GDPO. On LLaDA, it achieves excellent gains over LLaDA-Instruct, with an accuracy improvement of an average of **+2.40%** on GSM8K, **+3.00%** on MATH500, **+59.38%** on Countdown, and **+42.27%** on Sudoku. Compared with the strongest non-DMPO RL baseline at each setting, DMPO-LLaDA achieves especially large gains on planning tasks, including **+13.67%** and **+16.40%** on Countdown, and **+39.63%** and **+31.57%** on Sudoku at generation lengths 256 and 512, respectively. On Dream-Instruct (7B), DMPO-Dream also outperforms

Enhancing Reasoning for Diffusion LLMs via Distribution Matching Policy Optimization

Table 1. Model performances on reasoning benchmarks for LLaDA-Instruct (8B). **Best** and **second best** results are highlighted. DMPO consistently outperforms other baselines across different generation length.

Task Sequence Length	GSM8K			MATH500			Countdown			Sudoku		
	128	256	512	128	256	512	128	256	512	128	256	512
Dream-Instruct (7B)	56.63	73.39	76.65	31.00	36.60	36.40	22.66	28.52	27.34	14.45	16.41	11.77
LLaDA-Instruct (8B)	71.87	79.76	83.62	28.20	35.00	38.80	23.44	14.45	14.84	12.94	6.10	7.37
LLaDA-1.5 (8B)	73.09	80.97	84.38	26.80	33.80	40.00	26.17	16.41	23.83	15.19	13.04	8.98
d1-LLaDA	75.28	81.40	84.38	30.00	36.60	40.80	34.38	26.56	30.47	21.97	11.04	8.69
cGRPO-LLaDA	67.40	81.73	84.23	21.40	32.80	38.40	30.08	42.58	37.11	24.17	24.17	21.97
wd1-LLaDA	–	80.80	82.31	–	37.60	39.80	–	49.22	47.17	–	22.04	24.65
GDPO-LLaDA	–	81.20	82.26	–	38.00	38.20	–	67.19	66.41	–	24.17	25.10
DMPO-LLaDA (Ours)	74.83	82.41	85.22	30.00	38.20	42.80	67.19	80.86	82.81	32.76	63.80	56.67
DMPO-LLaDA-SFT (Ours)	80.06	84.00	84.09	31.80	40.00	41.20	54.69	67.19	77.34	25.20	25.73	23.78
DMPO-LLaDA-1.5 (Ours)	77.56	82.71	84.61	30.20	36.60	41.00	59.77	79.30	83.20	25.34	24.51	23.34

Table 2. Model performances on reasoning benchmarks for Dream-Instruct (7B). **Best** and **second best** results are highlighted. DMPO consistently outperforms other baselines across different generation length.

Task Sequence Length	GSM8K		MATH500		Countdown		Sudoku	
	256	512	256	512	256	512	256	512
Dream-Instruct (7B)	73.39	76.65	36.60	36.40	28.52	27.34	16.41	11.77
d1-Dream [†]	80.52	82.41	41.20	45.60	29.46	36.83	21.78	23.34
DMPO-Dream (Ours)	84.03	84.76	47.40	47.20	54.43	56.51	41.04	45.55

d1-Dream across all tasks, with average gains of **+2.93%** on GSM8K, **+3.90%** on MATH500, **+22.32%** on Countdown, and **+20.73%** on Sudoku. This underscores the overall effectiveness of DMPO for enhancing model reasoning capabilities across different dLLM backbones.

Weight baseline subtraction is crucial for small batch size training We test different choices for negative gradient insertion in Secs. 3.3 and 3.4 when training on the Sudoku dataset with a small batch size, and the results are visualized in Fig. 4. As shown by the curves, without weight baseline subtraction, the model does not improve as training progresses. All the proposed weight baselines in (13) to (15) effectively increase the reward value during training. WDDO achieves the fastest reward increase during the initial 1k steps but suffers from instability afterwards.

DMPO benefits from other means of post-training techniques To showcase the robustness and general efficacy of DMPO, we apply it to LLaDA-SFT and LLaDA-1.5. LLaDA-SFT is obtained by performing SFT of LLaDA-Instruct (8B) on s1k (Muennighoff et al., 2025), a dataset of 1k examples of high-quality reasoning questions with distilled reasoning traces from Gemini Thinking; LLaDA-

1.5 is obtained by performing DPO on 350K preference pairs covering a wide range of topics such as writing and reasoning. We then apply DMPO to these base models to obtain DMPO-LLaDA-SFT and DMPO-LLaDA-1.5, with performance reported in Tab. 1. DMPO continues to deliver performance gains for post-trained models, with consistent and significant accuracy improvements over base models, especially at generation lengths of 128 and 256 for the math reasoning datasets, with **+4.78%** and **+2.60%** on GSM8K, **+1.80%** and **+3.40%** on MATH500 compared with d1. This underscores that DMPO is a powerful method that integrates smoothly with existing solutions.

DMPO enables efficient and fast training Due to its *off-policy* and *forward* nature, DMPO achieves considerable training acceleration compared with GRPO-type methods. In Fig. 5, we compare head-to-head the training dynamics of diffu-GRPO, DMPO with ROAR sampler, and DMPO with Fast-dLLM (an approximate KV-cache mechanism enabled, confidence-based heuristic sampler for dLLMs, from Wu et al. (2026)) on Countdown under the same amount of training compute. Due to its off-policy nature, DMPO enables heavy reuse of each sampled buffer of rollouts and achieves sample efficiency $2 \sim 3 \times$ higher than that of diffu-GRPO.

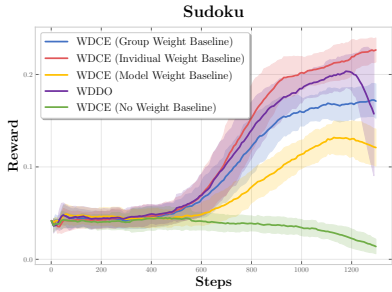


Figure 4. Effects of negative gradient insertion on Sudoku.

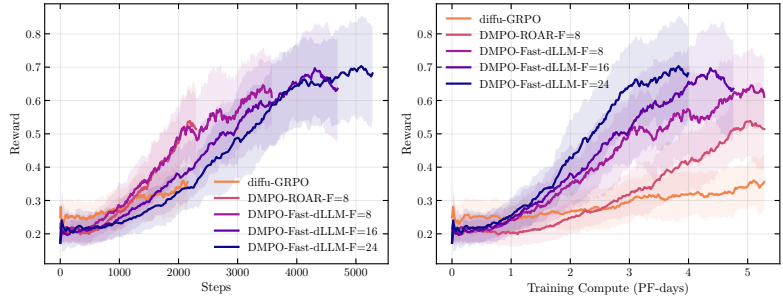


Figure 5. Comparison of training dynamics on Countdown. F is the frequency of sampling the buffer.

Regarding training-compute efficiency, as a forward-loss-based algorithm, DMPO offers flexibility in the choice of rollout sampler. With fast-dLLM, DMPO gains an acceleration of up to $8\times$ per rollout sampling, and achieves the same level of reward as d1 with only **31%** of the training budget (1.8 PF-days² v.s. 5.78 PF-days). This empirical evidence emphasizes that DMPO is not only an effective algorithm but also highly sample- and compute-efficient.

DMPO exhibits stable training despite highly stale data

As is evident from Figs. 5 and 6, DMPO enjoys a largely stable dynamics despite using up to $24\times$ stale data (which means 24 parameter updates on the same batch of rollouts), without suffering from high variance of importance sampling. While this seems to contradict the general belief that on-policy learning beats off-policy learning for LLM RL, we argue that this is *not* the case, because the off-policy in DMPO is inherently different from that used in diffu-GRPO or GRPO. Note that the latter considers the importance weight of the form $\frac{\pi_{\theta}}{\pi_{old}}$, which **inevitably diverges** as the number of parameter updates on θ increases. However, DMPO uses importance weight of the form $\frac{P_{\theta}}{P_{old}}$, which is independent of the current policy model π_{θ} and remains stable over a long horizon of training, enabling the use of a low buffer sampling frequency and highly stale rollouts without sacrificing performances. Moreover, DMPO adopts **sequence-level importance sampling**, in contrast to the *token-level* importance sampling used in diffu-GRPO or cGRPO, thereby providing an additional layer of stability. This advantage is also discussed in depth in Group Sequence Policy Optimization (GSPO, Zheng et al. (2025a)), which similarly considers *sequence-level* importance sampling.

Additional results and discussion can be found in App. C.3.

5. Conclusion

We propose Distribution Matching Policy Optimization (DMPO), a novel RL fine-tuning framework for dLLMs

that leverages their unique characteristics via importance sampling and a WDCE loss, enabling off-policy training and forward-only computation that naturally exploits dLLM inference capabilities. The main limitation of this work is that we focus on two pretrained dLLMs and four elementary reasoning benchmarks, and DMPO’s performance on other pretrained dLLMs and tasks in different domains remains unknown. Our work opens several directions for future research, such as investigating the distribution-matching framework for other generative models and studying the design of more effective weight-baseline techniques.

Acknowledgements

The authors are grateful for partial supports by NSF Grants ECCS-1942523, DMS-2206576, 2450378 (WG & YC), AFOSR Grant FA9550-25-1-0169 (WG & YC), Georgia Tech ARC-ACO Fellowship (WG), NSF Grant DMS-2513699 (YZ & MT), DOE Grants NA0004261 (MT), SC0026274 (YZ & MT), Richard Duke Fellowship (YZ & MT), and Simons Institute for the Theory of Computing at UC Berkeley (MT). This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. RRID:SCR.027619.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Anthropic. Introducing claude 4, May 2025. URL <https://www.anthropic.com/news/claude-4>. Accessed: 2025-09-01.

Arel. Arel’s sudoku generator. <https://www.ocf.be>

²1 PF-day = 8.64×10^{19} floating point operations.

- rkeley.edu/~arel/sudoku/main.html, 2025. Accessed: 2025-07-01.
- Arriola, M., Sahoo, S. S., Gokaslan, A., Yang, Z., Qi, Z., Han, J., Chiu, J. T., and Kuleshov, V. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tyEyYT267x>.
- Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=h7-XixPCAL>.
- Bai, J., Ye, T., Chow, W., Song, E., Chen, Q.-G., Li, X., Dong, Z., Zhu, L., and YAN, S. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=GJsuYHhAga>.
- Ben-Hamu, H., Gat, I., Severo, D., Nolte, N., and Karrer, B. Accelerated sampling from masked diffusion models via entropy bounded unmasking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=WBcBhTlNKO>.
- Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27381–27394. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e614f646836aaed9f89ce58e837e2310-Paper.pdf.
- Besnier, V., Chen, M., Hurych, D., Valle, E., and Cord, M. Halton scheduler for masked generative image transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=RDVrlWAb7K>.
- Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b5b528767aa35f5b1a60fe0aaeca0563-Paper-Conference.pdf.
- Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=kQwSbv0BR4>.
- Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. MaskGIT: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11315–11325, 2022. doi: 10.1109/CVPR52688.2022.01103.
- Chao, C.-H., Sun, W.-F., Liang, H., Lee, C.-Y., and Krishnan, R. Beyond masked and unmasked: Discrete diffusion models via partial masking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=nqpbvEZwF>.
- Chen, H., Ren, Y., Min, M. R., Ying, L., and Izzo, Z. Solving inverse problems via diffusion-based priors: An approximation-free ensemble sampling approach. *arXiv preprint arXiv:2506.03979*, 2025a.
- Chen, T., Zhang, Y., Tang, S., and Chatterjee, P. Multi-objective-guided discrete flow matching for controllable biological sequence design. *arXiv preprint arXiv:2505.07086*, 2025b.
- Chu, X., Huang, H., Zhang, X., Wei, F., and Wang, Y. GPG: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=mZn2Xyh9Ec>.
- DeepMind. Gemini diffusion. <https://deepmind.google/models/gemini-diffusion/>. Accessed: 2025-09-24.
- Deng, W., Ren, Y., Li, M., Sutherland, D. J., Li, X., and Thrampoulidis, C. On the effect of negative gradient in group relative deep reinforcement optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=2K9QsDaqkM>.

- Deschenaux, J. and Gulcehre, C. Beyond autoregression: Fast LLMs via self-distillation through time. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=uZ5K4HeNwd>.
- Domingo-Enrich, C., Han, J., Amos, B., Bruna, J., and Chen, R. T. Q. Stochastic optimal control matching. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 112459–112504. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cc32ec39a5073f61d38c338d963df30d-Paper-Conference.pdf.
- Domingo-Enrich, C., Drozdal, M., Karrer, B., and Chen, R. T. Q. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=xQBRrtQM8u>.
- Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/esser24a.html>.
- Gong, S., Zhang, R., Zheng, H., Gu, J., Jaitly, N., Kong, L., and Zhang, Y. DiffuCoder: Understanding and improving masked diffusion models for code generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=58NA3unZj5>.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.
- Guo, W., Choi, J., Zhu, Y., Tao, M., and Chen, Y. Proximal diffusion neural sampler. *arXiv preprint arXiv:2510.03824*, 2025b.
- Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, pp. 850–858, 2025. doi: 10.1126/science.ads0018.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.
- Hong, F., Yu, G., Ye, Y., Huang, H., Zheng, H., Zhang, Y., Wang, Y., and Yao, J. Wide-in, narrow-out: Revokable decoding for efficient and effective DLLMs. *arXiv preprint arXiv:2507.18578*, 2025.
- Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Hu, Z., Meng, J., Akhauri, Y., Abdelfattah, M. S., Seo, J.-s., Zhang, Z., and Gupta, U. Accelerating diffusion language model inference via efficient KV caching and guided diffusion. *arXiv preprint arXiv:2505.21467*, 2025.
- Inception Labs, Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 2025.
- Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Karimi Monsefi, A., Bhendawade, N., Ciosici, M. R., Culver, D., Zhang, Y., and Belousova, I. FS-DFM: Fast and accurate long text generation with few-step diffusion language models. *arXiv preprint arXiv:2509.20624*, 2025.
- Kim, J., Shah, K., Kontonis, V., Kakade, S. M., and Chen, S. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=DjJmre5IkP>.

- Kimi Team, Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., Li, C., Xiao, C., Du, C., Liao, C., et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.
- Li, X., Zhao, Y., Wang, C., Scalia, G., Eraslan, G., Nair, S., Biancalani, T., Ji, S., Regev, A., Levine, S., and Uehara, M. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=6QbbaEGk07>.
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.
- Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., ZHANG, D., and Ouyang, W. Flow-GRPO: Training flow matching models via online RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL <https://openreview.net/forum?id=oCBKGw5HNf>.
- Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.
- Liu, Z., Yang, Y., Zhang, Y., Chen, J., Zou, C., Wei, Q., Wang, S., and Zhang, L. dLLM-Cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*, 2025c.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/lou24a.html>.
- Ma, X., Yu, R., Fang, G., and Wang, X. dKV-Cache: The cache for diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=Gppo2JImHs>.
- Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- Nie, S., Zhu, F., Du, C., Pang, T., Liu, Q., Zeng, G., Lin, M., and Li, C. Scaling up masked diffusion models on text. 2025a. URL <https://openreview.net/forum?id=WNvvwK0tut>.
- Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL <https://openreview.net/forum?id=KnqiC0znVF>.
- Nisonoff, H., Xiong, J., Allenspach, S., and Listgarten, J. Unlocking guidance for discrete state-space diffusion and flow models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=XsgHl54y07>.
- Novikov, A., Vū, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J., Mehrabian, A., et al. AlphaEvolv: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.
- Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., and Li, C. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=sMyXP8Tanm>.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- Pan, J., Zhang, J., Wang, X., Yuan, L., Peng, H., and Suhr, A. Tinyzero. <https://github.com/Jiayi-Pan/TinyZero>, 2025. Accessed: 2025-01-24.
- Rector-Brooks, J., Hasan, M., Peng, Z., Liu, C.-H., Mittal, S., Dziri, N., Bronstein, M. M., Chatterjee, P., Tong, A., and Bose, J. Steering masked discrete diffusion models via discrete denoising posterior prediction. In *The*

- Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=Ombm8S40zn>.
- Ren, Y. and Sutherland, D. J. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tPNHOoZF19>.
- Ren, Y., Chen, H., Rotskoff, G. M., and Ying, L. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://openreview.net/forum?id=6awxwQE182>.
- Ren, Y., Chen, H., Zhu, Y., Guo, W., Chen, Y., Rotskoff, G. M., Tao, M., and Ying, L. Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL <https://openreview.net/forum?id=OuklL6Q3s0>.
- Ren, Y., Gao, W., Ying, L., Rotskoff, G. M., and Han, J. DriftLite: Lightweight drift control for inference-time scaling of diffusion models. *arXiv preprint arXiv:2509.21655*, 2025c.
- Rojas, K., Lin, J., Rasul, K., Schneider, A., Nevmyvaka, Y., Tao, M., and Deng, W. Improving reasoning for diffusion language models via group diffusion policy optimization. *arXiv preprint arXiv:2510.08554*, 2025a.
- Rojas, K., Zhu, Y., Zhu, S., Ye, F. X.-F., and Tao, M. Diffuse everything: Multimodal diffusion models on arbitrary state spaces. In *Forty-second International Conference on Machine Learning*, 2025b. URL <https://openreview.net/forum?id=AjbiIcRt6q>.
- Rojas, K., He, Y., Lai, C.-H., Takida, Y., Mitsuji, Y., and Tao, M. Improving classifier-free guidance in masked diffusion: Low-dim theoretical insights with high-dim impact. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=mMK9pvQJxf>.
- Sahoo, S. S., Arriola, M., Gokaslan, A., Marroquin, E. M., Rush, A. M., Schiff, Y., Chiu, J. T., and Kuleshov, V. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=L4uaAR4ArM>.
- Sahoo, S. S., Yang, Z., Akhauri, Y., Liu, J., Singh, D., Cheng, Z., Liu, Z., Xing, E., Thickstun, J., and Vahdat, A. Esoteric language models. *arXiv preprint arXiv:2506.01928*, 2025.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/schulman15.html>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=xcqSOfHt4g>.
- Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Song, S., Tong, Y., Li, X., Li, X., et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.
- Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.
- Tang, S., Zhang, Y., and Chatterjee, P. PepTune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025a. URL <https://openreview.net/forum?id=FQoy1Y1Hd8>.
- Tang, S., Zhu, Y., Tao, M., and Chatterjee, P. TR2-D2: Tree search guided trajectory-aware fine-tuning for discrete diffusion. *arXiv preprint arXiv:2509.25171*, 2025b.
- Tang, X., Dolga, R., Yoon, S., and Bogunovic, I. wd1: Weighted policy optimization for reasoning in diffusion language models. *arXiv preprint arXiv:2507.08838*, 2025c.
- Uria, B., Côté, M.-A., Gregor, K., Murray, I., and Larochelle, H. Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17(205): 1–37, 2016. URL <http://jmlr.org/papers/v17/16-272.html>.
- von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrusch, T., Lambert, N., Huang, S., Rasul, K., and Galouédec, Q. TRL: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.

- Wang, C., Uehara, M., He, Y., Wang, A., Lal, A., Jaakkola, T., Levine, S., Regev, A., Hanchen, and Biancalani, T. Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://openreview.net/forum?id=G328D1xt4W>.
- Wang, Y., Yang, L., Li, B., Tian, Y., Shen, K., and Wang, M. Revolutionizing reinforcement learning framework for diffusion large language models. *arXiv preprint arXiv:2509.06949*, 2025b.
- Weng, L. Reward hacking in reinforcement learning. *lilianweng.github.io*, Nov 2024. URL <https://lilianweng.github.io/posts/2024-11-28-reward-hacking/>.
- Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., and Xie, E. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=3Z3Is6hnOT>.
- Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.
- Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., and Wang, M. MMaDA: Multimodal large diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=wczmXLuLGd>.
- Yao, F., Liu, L., Zhang, D., Dong, C., Shang, J., and Gao, J. Your efficient RL framework secretly brings you off-policy RL training, August 2025. URL <https://fengyao.notion.site/off-policy-rl>.
- Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., and Kong, L. Dream 7B: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.
- Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., YuYue, Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Yan, L., Wu, Y., and Wang, M. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=2a36EMSSTp>.
- Zekri, O. and Boullé, N. Fine-tuning discrete diffusion models with policy gradient methods. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=rXFzVRZsbt>.
- Zhang, R., Zhai, S., Gu, J., Zhang, Y., Zheng, H., Chen, T., Bautista, M. Á., Susskind, J. M., and Jaitly, N. Flexible language modeling in continuous space with transformer-based autoregressive flows. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL <https://openreview.net/forum?id=MR7Fn23hSE>.
- Zhang, R., Zhai, S., Zhang, Y., Thornton, J., Ou, Z., Susskind, J. M., and Jaitly, N. Target concrete score matching: A holistic framework for discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025b. URL <https://openreview.net/forum?id=ZMrdvSm7xi>.
- Zhao, S., Gupta, D., Zheng, Q., and Grover, A. d1: Scaling reasoning in diffusion large language models via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL <https://openreview.net/forum?id=7ZVRlBFuEv>.
- Zhao, Y., Liu, Y., Liu, J., Chen, J., Wu, X., Hao, Y., Lv, T., Huang, S., Cui, L., Ye, Q., et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025b.
- Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., Zhou, J., and Lin, J. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.
- Zheng, H., Gong, S., Zhang, R., Chen, T., Gu, J., Zhou, M., Jaitly, N., and Zhang, Y. Continuously augmented discrete diffusion model for categorical generative modeling. *arXiv preprint arXiv:2510.01329*, 2025b.
- Zheng, H., Liu, X., Kong, C. X., Jiang, N., Hu, Z., Luo, W., Deng, W., and Lin, G. Ultra-fast language generation via discrete diffusion divergence instruct. *arXiv preprint arXiv:2509.25035*, 2025c.
- Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., and Liu, M.-Y. DiffusionNFT: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025d.
- Zheng, K., Chen, Y., Chen, H., He, G., Liu, M.-Y., Zhu, J., and Zhang, Q. Direct discriminative optimization: Your likelihood-based visual generative model is secretly a GAN discriminator. In *Forty-second International Conference on Machine Learning*, 2025e. URL <https://openreview.net/forum?id=OJ6WE7F8tK>.

- Zheng, K., Chen, Y., Mao, H., Liu, M.-Y., Zhu, J., and Zhang, Q. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *The Thirteenth International Conference on Learning Representations*, 2025f. URL <https://openreview.net/forum?id=CTC7CmirNr>.
- Zhou, C., Yang, C., Hu, Y., Wang, C., Zhang, C., Zhang, M., Mackey, L., Jaakkola, T., Bates, S., and Zhang, D. Coevolutionary continuous discrete diffusion: Make your diffusion language model a latent reasoner. *arXiv preprint arXiv:2510.03206*, 2025.
- Zhu, F., Wang, R., Nie, S., Zhang, X., Wu, C., Hu, J., Zhou, J., Chen, J., Lin, Y., Wen, J.-R., et al. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*, 2025a.
- Zhu, S., Zhu, Y., Tao, M., and Qiu, P. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=FtjLUHyZAO>.
- Zhu, X., Cheng, D., Zhang, D., Li, H., Zhang, K., Jiang, C., Sun, Y., Hua, E., Zuo, Y., Lv, X., et al. FlowRL: Matching reward distributions for LLM reasoning. *arXiv preprint arXiv:2509.15207*, 2025c.
- Zhu, Y., Chen, T., Kong, L., Theodorou, E., and Tao, M. Trivialized momentum facilitates diffusion generative modeling on Lie groups. In *The Thirteenth International Conference on Learning Representations*, 2025d. URL <https://openreview.net/forum?id=DTatjJTD11>.
- Zhu, Y., Guo, W., Choi, J., Liu, G.-H., Chen, Y., and Tao, M. MDNS: Masked diffusion neural sampler via stochastic optimal control. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025e. URL <https://openreview.net/forum?id=xIH95kXNR2>.
- Zhu, Y., Wang, X., Lathuilière, S., and Kalogeiton, V. Di[M]O: Distilling masked diffusion models into one-step generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18606–18618, October 2025f.
- Zhu, Y., Wang, X., Lathuilière, S., and Kalogeiton, V. Soft-Di[M]O: Improving one-step discrete image generation with soft embeddings. *arXiv preprint arXiv:2509.22925*, 2025g.

Algorithm 1 Distribution Matching Policy Optimization

Require: Training dataset \mathcal{D} , number of prompts per batch B , number of rollouts per prompt N , frequency for sampling buffer F , model policy π_θ .

- 1: **for** step = 0, 1, 2, ... **do**
- 2: **if** step mod $F = 0$ **then**
- 3: {Prepare the buffer using the current policy, denoted π_θ .}
- 4: Sample B prompts $\{\mathbf{q}^{(i)}\}_{1 \leq i \leq B}$ from the dataset \mathcal{D} .
- 5: **for** $1 \leq i \leq B$ (in parallel, with gradient computation disabled) **do**
- 6: Sample N orders and generate N rollouts $\{\mathbf{o}^{(i,n)}\}_{1 \leq n \leq N}$ conditional on prompt $\mathbf{q}^{(i)}$.
- 7: Evaluate reward and compute weights $w(\mathbf{o}^{(i,n)}|\mathbf{q}^{(i)}; \boldsymbol{\sigma}^{(i,n)})$ according to (10).
- 8: Compute the weight baseline according to (13), (14), or (15), and obtain the real weights $w_{\text{real}}(\mathbf{o}^{(i,n)}|\mathbf{q}^{(i)}; \boldsymbol{\sigma}^{(i,n)})$ according to (12).
- 9: **end for**
- 10: **end if**
- 11: For each $\mathbf{o}^{(i,n)}$, sample a mask assignment and obtain $\tilde{\mathbf{o}}^{(i,n)}$.
- 12: Feed all pairs of $(\mathbf{q}^{(i)}, \tilde{\mathbf{o}}^{(i,n)})$ into π_θ and compute the WDCE loss (11), then update θ .
- 13: **end for**

output π_θ

A. Related Work

Here, we focus on the literature for discrete diffusion models, as well as the methods for fine-tuning MDMs, dLLMs, and LLMs. We also briefly review several GRPO-style algorithms for domains outside of LLMs.

Discrete diffusion models Diffusion models have been top-performing approaches for generating various data modalities (Zhu et al., 2025d; Esser et al., 2024; Zhu et al., 2025b; Rojas et al., 2025b; Zheng et al., 2025e; Chen et al., 2025a; Ren et al., 2025c). Discrete diffusion models (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024; Zhang et al., 2025b), a natural extension of diffusion models to finite state spaces, have emerged as powerful approaches for generating categorical, sequence data, with applications to text (Nie et al., 2025a;b; Ye et al., 2025), images (Chang et al., 2022; Bai et al., 2025; Shi et al., 2025), and biological sequences (Tang et al., 2025a; Chen et al., 2025b). One of the most effective variants of discrete diffusion models is masked diffusion models (MDM) (Sahoo et al., 2024; Ou et al., 2025; Shi et al., 2024) and its variants (Arriola et al., 2025; Sahoo et al., 2025; Chao et al., 2025). Recently, continuous latents have also been introduced into the modeling of discrete data (Zhang et al., 2025a; Zhou et al., 2025; Zheng et al., 2025b), resulting in improved and more appealing performance.

One particularly important line of development for discrete diffusion models centers on their inference techniques, with the aim of improving generation quality (Nisonoff et al., 2025; Rojas et al., 2026; Kim et al., 2025; Besnier et al., 2025) and accelerating sampling speed (Ren et al., 2025b; Ben-Hamu et al., 2025; Wu et al., 2026; Hong et al., 2025). Besides these training-free approaches, learning-based approaches, such as few-step distillation, have also achieved decent success for discrete diffusion models (Deschenaux & Gulcehre, 2025; Karimi Monsefi et al., 2025; Zheng et al., 2025c; Zhu et al., 2025f;g). DMPO is closely tied to the literature on fast inference, as it can benefit from it by enjoying a similar training speed acceleration due to its forward nature.

Fine-tuning general discrete diffusion models Earlier works on fine-tuning discrete diffusion models primarily focus on applications in biological and chemical domains, e.g., SVDD (Li et al., 2025), DDPP (Rector-Brooks et al., 2025), DRAKES (Wang et al., 2025a), SEPO (Zekri & Boullé, 2025), and TR2-D2 (Tang et al., 2025b). Although these methods work well for their respective tasks, they are not directly applicable to dLLMs due to the unique challenges posed by the language domain, such as large model size, high dimensionality, and the need to maintain linguistic coherence and diversity.

Fine-tuning diffusion LLMs Recently, numerous works have proposed RL algorithms for fine-tuning dLLMs, with most existing works being adaptations of the GRPO algorithm (Shao et al., 2024) for AR LLMs. For example, Zhao et al. (2025a) proposed Diffu-GRPO that estimates the per-token response log probabilities via masking all except the required response positions, and partially masking the prompt to get the model output, while their sequence log probability is estimated

by mean-field approximation. Gong et al. (2026) introduced Coupled GRPO that modified the Diffu-GRPO method by not partially masking the prompt, and using complementary pairs of masks to mask the same response that fully uses the model output, which we also adopt in our experiments. Yang et al. (2025) proposed UniGRPO, which involves a structured noise strategy and a modified log-likelihood approximation (both per-token and sequence). Concurrent with our work, TraceRL (Wang et al., 2025b) improves dLLM RL training by minimizing a training-inference gap. wd1 (Tang et al., 2025c) introduces additional regularization to the old policy, alongside the regularization applied to the reference model policy, which resembles the case discussed in App. B.2. We highlight that all these methods are GRPO-style algorithms that require estimating per-token response log probabilities, which are typically intractable and challenging for dLLMs. In contrast, our method offers the advantage of being a forward one, with greater efficiency and accuracy.

Fine-tuning LLMs For fine-tuning LLMs, pre-LLM era works such as Trust Region Policy Optimization (TRPO, Schulman et al. (2015)) and Proximal Policy Optimization (PPO, Schulman et al. (2017)) have been widely used for RLHF (Ouyang et al., 2022). Since the huge success of GRPO (Shao et al., 2024) on DeepSeek-R1 (Guo et al., 2025a), there have been many follow-up works that improve GRPO in various ways, for instance: GRPO Done Right (Dr-GRPO, Liu et al. (2025b)), Decoupled clip and Dynamic sAmpling Policy Optimization (DAPO, Yu et al. (2025)), Group Policy Gradient (GPG, Chu et al. (2025)), Group Sequence Policy Optimization (GSPO, Zheng et al. (2025a)), Geometric-Mean Policy Optimization (GMPO, Zhao et al. (2025b)), etc.

Apart from the aforementioned policy gradient-based methods, GFlowNet (Bengio et al., 2021) has also been applied to finetuning LLMs, with successful applications seen in Kimi 1.5 (Kimi Team et al., 2025) and FlowRL (Zhu et al., 2025c). Notably, concurrent with our work, FlowRL shares the same high-level goal as our DMPO, targeting also policy distribution matching rather than merely reward maximization for AR-LLMs. However, distinct from DMPO, FlowRL derives its objectives from reverse KL and utilizes GFlowNet objectives. In contrast, our approach considers forward KL, which is known to be mass-covering, and implements it using importance sampling and weighted denoising cross-entropy.

GRPO-style algorithms for fine-tuning diffusion and flow-based models GRPO-type algorithms have also been adapted to diffusion and flow-based models, such as flow-GRPO (Liu et al., 2025a) and DanceGRPO (Xue et al., 2025). Aside from that, there are also SOC-based fine-tuning algorithms for diffusion models, such as adjoint matching (Domingo-Enrich et al., 2025), with which our work shares similarity at a high level. Concurrent with our work, DiffusionNFT (Zheng et al., 2025d) has been proposed to finetune continuous diffusion models for text-to-image generation tasks. While formulated in drastically different ways, DiffusionNFT shares a similarity with our DMPO in that it is also an algorithm that primarily depends on model forward passes rather than backward trajectories.

B. Theory of Distribution Matching Policy Optimization

B.1. Distribution Matching Policy Optimization from the Stochastic Optimal Control Perspective

This section aims at providing an alternative derivation of DMPO from the perspective of stochastic optimal control (SOC), which is inspired by DRAKES (Wang et al., 2025a) and MDNS (Zhu et al., 2025e). We will first introduce the necessary background on continuous-time Markov chains (CTMCs), then show how MDM sampling can be viewed as a CTMC. Finally, we derive the DMPO framework from the SOC perspective.

Introduction to Continuous-time Markov Chains To derive the SOC framework for fine-tuning, we view the sampling of an MDM as a time-indexed stochastic process, and the proper mathematical tool is the **continuous-time Markov chain (CTMC)**. A CTMC $X = (X_t)_{t \in [0,1]}$ is a stochastic process taking value in a discrete state space \mathcal{X} . Its law is characterized by the **rate matrix** $Q = (Q_t)_{t \in [0,1]}$, defined as

$$Q_t(x, y) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(X_{t+\Delta t} = y | X_t = x) - 1_{x=y}}{\Delta t}, \forall x, y \in \mathcal{X}. \quad (17)$$

By definition, the off-diagonal entries of Q_t are non-negative, and each row sums to zero.

The **path** of X , i.e., $t \mapsto X_t(\omega)$, is piecewise constant with discontinuous jumps, and one typically assumes that the path is right continuous with left limits. The **path measure** a CTMC X is a probability measure on the space of paths defined as $\mathbb{P}^X(A) := \Pr(X \in A)$, which is the distribution of X . The following lemma shows how to compute the **Radon-Nikodým (RN) derivative** between two path measures driven by CTMCs with different rate matrices and initial distributions:

Lemma B.1. Given two CTMCs with rate matrices Q^1, Q^2 and initial distributions μ_1, μ_2 on \mathcal{X} , let $\mathbb{P}^1, \mathbb{P}^2$ be the associated path measures. Then, for any path $\xi = (\xi_t)_{t \in [0,1]}$,

$$\log \frac{d\mathbb{P}^1}{d\mathbb{P}^2}(\xi) = \log \frac{d\mu_1}{d\mu_2}(\xi_0) + \sum_{t: \xi_{t-} \neq \xi_t} \log \frac{Q_t^1(\xi_{t-}, \xi_t)}{Q_t^2(\xi_{t-}, \xi_t)} + \int_0^1 (Q_t^1(\xi_t, \xi_t) - Q_t^2(\xi_t, \xi_t)) dt. \quad (18)$$

For the proof, see [Campbell et al. \(2024, App. C.1\)](#), [Ren et al. \(2025a, Thm. 3.3\)](#), or [Zhu et al. \(2025e, Lem. 1\)](#). An intuitive interpretation of (18) is to view the RN derivative as the limit of density ratios between finite-dimensional joint distributions, and approximate the transition probability by (17).

Masked Diffusion Models as Continuous-Time Markov Chains We will now delve into the CTMC formulation of sampling from an MDM. To avoid notational clutter, we use superscript to denote the position index, and subscript to denote the time index (e.g., $\xi_t = (\xi_t^1, \dots, \xi_t^D)$). We present the theory only in the case of *unconditional generation* with sequence length D for simplicity of notations, but it can be easily easily generalized to the case of conditional generation of \mathbf{o} given a prompt \mathbf{q} .

As shown in ([Ou et al., 2025](#)), by introducing a noise schedule $\gamma(t) = \frac{1}{t}$,³ the random order autoregressive sampling of an MDM π_θ can be viewed as a CTMC with the rate matrix $Q^\theta = (Q_t^\theta)_{t \in [0,1]}$ such that for $\mathbf{x} \neq \mathbf{y} \in \bar{\mathcal{V}}^D$,

$$Q_t^\theta(\mathbf{x}, \mathbf{y}) = \gamma(t) \pi_\theta(\mathbf{x})_{d,n}, \text{ if } \mathbf{x}^d = \mathbf{M} \text{ and } \mathbf{y} = \mathbf{x}^{d \leftarrow n},$$

and 0 if otherwise, where $\mathbf{x}^{d \leftarrow n}$ means the sequence obtained by replacing the d -th position of \mathbf{x} by n . The diagonal terms of Q_t^θ can be computed as

$$\begin{aligned} Q_t^\theta(\mathbf{x}, \mathbf{x}) &= - \sum_{\mathbf{y} \neq \mathbf{x}} Q_t^\theta(\mathbf{x}, \mathbf{y}) = - \sum_{d: \mathbf{x}^d = \mathbf{M}} \sum_n Q_t^\theta(\mathbf{x}, \mathbf{x}^{d \leftarrow n}) \\ &= - \sum_{d: \mathbf{x}^d = \mathbf{M}} \sum_n \gamma(t) \pi_\theta(\mathbf{x})_{d,n} = -\gamma(t) \cdot |\{d : \mathbf{x}^d = \mathbf{M}\}|. \end{aligned} \quad (19)$$

Therefore, if $\mathbb{P}^\theta, \mathbb{P}^{\theta'}$ are the path measures of the sampling processes of two MDMs parameterized by θ and θ' , respectively, then by (18), assuming that the jump from ξ_{t-} to ξ_t is at the $d(t)$ -th position, we have

$$\log \frac{d\mathbb{P}^{\theta'}}{d\mathbb{P}^\theta}(\xi) = \sum_{t: \xi_{t-} \neq \xi_t} \log \frac{\pi_{\theta'}(\xi_{t-})_{d(t), \xi_t^{d(t)}}}{\pi_\theta(\xi_{t-})_{d(t), \xi_t^{d(t)}}}, \forall \xi = (\xi_t)_{t \in [0,1]}, \quad (20)$$

as the first term in (18) is always zero (both initial distributions are the point mass on the fully masked sequence), and the diagonal terms in the third term cancel out due to (19).

Moreover, as proved in [Ou et al. \(2025\)](#), the training of an MDM π_θ given i.i.d. samples from the target distribution p_{data} can be interpreted as minimizing the KL divergence between the target path measure \mathbb{P}^* and the parameterized path measure \mathbb{P}^θ , where \mathbb{P}^* is defined as the path measure of the CTMC with rate matrix $Q_t^*(\mathbf{x}, \mathbf{x}^{d \leftarrow n}) = \gamma(t) \Pr_{\mathbf{X} \sim p_{\text{data}}}(X^d = n | \mathbf{X}^{\text{UM}} = \mathbf{x}^{\text{UM}})_{1_{x^d = \mathbf{M}}}$, i.e., with the ground-truth conditional distribution. Moreover, one can derive

$$\text{KL}(\mathbb{P}^* || \mathbb{P}^\theta) = \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathbb{E}_{m \sim \text{Unif}\{1, \dots, D\}} \left[\frac{D}{m} \mathbb{E}_{\mu_m(\tilde{\mathbf{x}}|\mathbf{x})} \sum_{d: \tilde{\mathbf{x}}^d = \mathbf{M}} -\log \pi_\theta(\tilde{\mathbf{x}})_{d, x^d} \right] + \text{const},$$

where const does not depend on θ , and $\mu_m(\cdot|\mathbf{x})$ means to sample a uniformly random subset of $\{1, \dots, D\}$ of size m and mask the corresponding positions in \mathbf{x} . Note that this is exactly the denoising cross-entropy loss $\mathbb{E}_{p_{\text{data}}(\mathbf{x})} \mathcal{L}_\theta(\mathbf{x})$ as presented in (2). In other words, minimizing the KL divergence between sequence-level probabilities $p_*(\xi_1)$ and $p_\theta(\xi_1) \approx e^{-\mathcal{L}_\theta(\xi_1)}$ in (11) can be interpreted as *precisely* minimizing the KL divergence between path-level probabilities $\mathbb{P}^*(\xi)$ and $\mathbb{P}^\theta(\xi)$.

³The choice of noise schedule is essentially not important for MDM. In fact, γ can be any positive function with $\int_0^1 \gamma(t) dt = \infty$. Here, we follow the convention in most of the literature on MDM and choose this specific γ such that the conditional distribution of $\xi_t \in \bar{\mathcal{V}}^D$ given $\xi_1 \in \mathcal{V}^D$ is obtained by independently masking each position in ξ_1 with probability $1 - t$.

Fine-tuning MDMs as a Stochastic Optimal Control Problem on Path Measures The task of fine-tuning a pretrained MDM can be viewed as a stochastic optimal control (SOC) problem on the space of path measures: given a pretrained MDM π_{ref} which generates a distribution p_{ref} , we define its induced **reference path measure** as \mathbb{P}^{ref} , with rate matrix $Q_t^{\text{ref}}(\mathbf{x}, \mathbf{x}^{d \leftarrow n}) = \gamma(t)\pi_{\text{ref}}(\mathbf{x})_{d,n} \mathbf{1}_{x^d = \mathbf{M}}$, and has terminal distribution $\mathbb{P}_1^{\text{ref}} = p_{\text{ref}}$. We aim at finding a target rate matrix Q^* such that the associated target path measure \mathbb{P}^* has a terminal distribution p_* defined in the following way of tilting by reward:

$$p_*(\mathbf{x}) = \frac{1}{Z} p_{\text{ref}}(\mathbf{x}) e^{r(\mathbf{x})/\alpha}, \quad \mathbf{x} \in \mathcal{V}^D, \quad \text{where } Z = \sum_{\mathbf{x}} p_{\text{ref}}(\mathbf{x}) e^{r(\mathbf{x})/\alpha}.$$

This can be achieved by defining the target path measure \mathbb{P}^* as

$$\mathbb{P}^*(\xi) = \mathbb{P}^{\text{ref}}(\xi_{[0,1]} | \xi_1) p_*(\xi_1) = \mathbb{P}^{\text{ref}}(\xi) \frac{p_*(\xi_1)}{p_{\text{ref}}(\xi_1)} = \frac{1}{Z} \mathbb{P}^{\text{ref}}(\xi) e^{r(\xi_1)/\alpha}, \quad \forall \xi = (\xi_t)_{t \in [0,1]}. \quad (21)$$

We use a network π_θ to parameterize the new rate matrix, initialized at π_{ref} . Given a current path measure \mathbb{P}^θ induced by a CTMC with rate matrix $Q_t^\theta(\mathbf{x}, \mathbf{x}^{d \leftarrow n}) = \gamma(t)\pi_\theta(\mathbf{x})_{d,n} \mathbf{1}_{x^d = \mathbf{M}}$, we can first derive the RN derivative between the path measures by (20):

$$\begin{aligned} \log \frac{d\mathbb{P}^*}{d\mathbb{P}^\theta}(\xi) &= \log \frac{d\mathbb{P}^*}{d\mathbb{P}^{\text{ref}}}(\xi) + \log \frac{d\mathbb{P}^\theta}{d\mathbb{P}^{\text{ref}}}(\xi) \\ &= \frac{r(\xi_1)}{\alpha} - \log Z + \sum_{t: \xi_{t-} \neq \xi_t} \log \frac{Q_t^\theta(\xi_{t-}, \xi_t)}{Q_t^{\text{ref}}(\xi_{t-}, \xi_t)} + \int_0^1 \sum_{y \neq \xi_t} (Q_t^\theta - Q_t^{\text{ref}})(\xi_t, y) dt \\ &= \frac{r(\xi_1)}{\alpha} + \sum_{t: \xi_{t-} \neq \xi_t} \log \frac{\pi_{\text{ref}}(\xi_{t-})_{d(t), \xi_t^{d(t)}}}{\pi_\theta(\xi_{t-})_{d(t), \xi_t^{d(t)}}} - \log Z =: W^\theta(\xi) - \log Z, \end{aligned} \quad (22)$$

where we assume that the jump from ξ_{t-} to ξ_t is at the $d(t)$ -th position. The idea of the weighted denoising cross-entropy (WDCE) loss is essentially to treat i.i.d. samples from the current policy \mathbb{P}^θ as weighted samples from \mathbb{P}^* , and minimizing the following loss:

$$\begin{aligned} \text{KL}(\mathbb{P}^* \| \mathbb{P}^\theta) + \text{const} &= \mathbb{E}_{p_*(\mathbf{x})} \mathcal{L}_\theta(\mathbf{x}) = \mathbb{E}_{\mathbb{P}^*(\xi)} \mathcal{L}_\theta(\xi_1) \\ &= \mathbb{E}_{\mathbb{P}^v(\xi)} \frac{d\mathbb{P}^*}{d\mathbb{P}^v}(\xi) \mathcal{L}_\theta(\xi_1) = \mathbb{E}_{\mathbb{P}^v(\xi)} \frac{1}{Z} e^{W^v(\xi)} \mathcal{L}_\theta(\xi_1), \end{aligned}$$

where \mathbb{P}^v is the path measure induced by a CTMC with rate matrix Q^v where the network is parameterized by v (e.g., the old parameters θ_{old}), whose parameters do not involve gradient calculation. For instance, we can set $v = \theta_{\text{old}}$. Note that $Z = \mathbb{E}_{\mathbb{P}^v(\xi)} e^{W^v(\xi)}$, which, if estimated via samples, is equivalent to doing softmax normalization on the logits $W^v(\xi)$ in the batch. Comparing with the WDCE loss (11) presented in Sec. 3.2, we conclude that they are essentially the same.

B.2. Generalizing WDCE to Zero Temperature with Proximal Descent

Recall that our target distribution is (5), which is under a temperature $\alpha > 0$. We propose to generalize the WDCE loss (11) to incorporate the limiting case $\alpha \rightarrow 0$ from the viewpoint of **proximal descent** (Guo et al., 2025b).

The reward maximization problem (4) provides a variational characterization of the target distribution $p_*(\mathbf{o}|\mathbf{q})$. Suppose now we have a dLLM policy $\pi_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})$ that outputs a distribution $p_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})$. We define the next target distribution $p_{\text{tar}}(\mathbf{o}|\mathbf{q})$ as

$$p_{\text{tar}}(\mathbf{o}|\mathbf{q}) = \operatorname{argmax}_{p_\theta(\mathbf{o}|\mathbf{q})} \left\{ \mathbb{E}_{p_\theta(\mathbf{o}|\mathbf{q})} [r(\mathbf{q}, \mathbf{o})] - \alpha \text{KL}(p_\theta(\cdot|\mathbf{q}) \| p_{\text{ref}}(\cdot|\mathbf{q})) - \frac{1}{\eta'} \text{KL}(p_\theta(\cdot|\mathbf{q}) \| p_{\theta_{\text{old}}}(\cdot|\mathbf{q})) \right\}, \quad (23)$$

where $\eta' > 0$ is the step size. Let $\eta = \frac{\eta'}{1+\eta'/\alpha} \in (0, \frac{1}{\alpha})$. It is easy to see that the solution is given by

$$\begin{aligned} p_{\text{tar}}(\mathbf{o}|\mathbf{q}) &\propto_{\mathbf{o}} p_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})^{1-\eta\alpha} p_{\text{ref}}(\mathbf{o}|\mathbf{q})^{\eta\alpha} e^{\eta r(\mathbf{q}, \mathbf{o})}, \\ &\propto_{\mathbf{o}} p_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})^{1-\eta\alpha} p_*(\mathbf{o}|\mathbf{q})^{\eta\alpha}. \end{aligned} \quad (24)$$

In fact, the term inside the brackets in (23) is $-\frac{1}{\eta} \text{KL}(p_\theta(\cdot|\mathbf{q})\|p_{\text{tar}}(\cdot|\mathbf{q})) + \text{const}$. This means the next target distribution is a geometric interpolation between the current model distribution $p_{\theta_{\text{old}}}$ and the optimal distribution p_* , with $\eta > 0$ being a step size parameter. (24) is well-defined even when $\alpha = 0$, although in this case, the target distribution concentrates on the set of maximizers of $r(\mathbf{q}, \mathbf{o})$ (e.g., all correct question-response pairs) without regularization from the base model $p_{\text{ref}}(\mathbf{o}|\mathbf{q})$.

For $\alpha = 0$, $p_{\text{tar}}(\mathbf{o}|\mathbf{q}) \propto_{\mathbf{o}} p_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})e^{\eta r(\mathbf{q}, \mathbf{o})}$. We can similarly solve the distribution matching problem via the WDCE loss:

$$\begin{aligned} \text{KL}(p_{\text{tar}}(\cdot|\mathbf{q})\|p_\theta(\cdot|\mathbf{q})) &= \mathbb{E}_{p_{\text{tar}}(\mathbf{o}|\mathbf{q})}[-\log p_\theta(\mathbf{o}|\mathbf{q})] + \text{const} \\ &= \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q})} \underbrace{\frac{p_{\text{tar}}(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})}}_{=:w(\mathbf{o}|\mathbf{q})} [-\log p_\theta(\mathbf{o}|\mathbf{q})] + \text{const} \\ &\leq \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q})} w(\mathbf{o}|\mathbf{q}) \mathcal{L}_\theta(\mathbf{o}|\mathbf{q}) + \text{const}, \end{aligned}$$

where the importance weight $w(\mathbf{o}|\mathbf{q}) \propto_{\mathbf{o}} \exp\left(\eta r(\mathbf{q}, \mathbf{o}) + \log \frac{p_{\theta_{\text{old}}}(\mathbf{o}|\mathbf{q})}{p_{\theta_v}(\mathbf{o}|\mathbf{q})}\right)$. For $v \leftarrow \theta_{\text{old}}$, the weight simplifies to the softmax of $\eta r(\mathbf{q}, \mathbf{o})$ over all responses for the same prompt \mathbf{q} . The weight baseline subtraction tricks also apply here.

We remark that when picking $\alpha = 0$, through the proximal gradient descent formulation, DMPO becomes completely *forward-only*, as it eliminates the need for estimating the sequence log probability ratio of the form $\log \frac{p_{\text{ref}}(\mathbf{o}|\mathbf{q})}{p_v(\mathbf{o}|\mathbf{q})}$, making it the best option to incorporate fast dLLM inference techniques for RL training speed-up. However, in this case, we can no longer guarantee the diversity in the target optimal distribution, and thus, we save this direction for future investigation.

B.3. Insights for Weight Baselines: Approximate Variance Reduction

We first recall a classical equality in statistics regarding the **score function**: if $p_\theta(x)$ is a probability density or probability mass function parameterized by a continuous parameter θ , then under certain weak regularity conditions, we have $\mathbb{E}_{p_\theta(x)} \nabla_\theta \log p_\theta(x) = 0$.

Therefore,

$$\begin{aligned} 0 &= \mathbb{E}_{p_\theta(\mathbf{o}|\mathbf{q})} \nabla_\theta \log p_\theta(\mathbf{o}|\mathbf{q}) = \nabla_\theta \mathbb{E}_{p_{\bar{\theta}}(\mathbf{o}|\mathbf{q})} \log p_\theta(\mathbf{o}|\mathbf{q}) \\ &= \nabla_\theta \mathbb{E}_\sigma \mathbb{E}_{p_{\bar{\theta}}(\mathbf{o}|\mathbf{q}; \sigma)} \log p_\theta(\mathbf{o}|\mathbf{q}) \\ &= \nabla_\theta \mathbb{E}_\sigma \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q}; \sigma)} \frac{p_{\bar{\theta}}(\mathbf{o}|\mathbf{q}; \sigma)}{p_v(\mathbf{o}|\mathbf{q}; \sigma)} \log p_\theta(\mathbf{o}|\mathbf{q}). \end{aligned}$$

Combined with (9), we can see that subtracting $\frac{p_{\bar{\theta}}(\mathbf{o}|\mathbf{q}; \sigma)}{p_v(\mathbf{o}|\mathbf{q}; \sigma)}$ from the weight does not change the gradient of the CE loss, i.e.,

$$\nabla_\theta \text{KL}(p_*(\cdot|\mathbf{q})\|p_\theta(\cdot|\mathbf{q})) = \nabla_\theta \mathbb{E}_\sigma \mathbb{E}_{p_v(\mathbf{o}|\mathbf{q}; \sigma)} \left(\frac{p_*(\mathbf{o}|\mathbf{q}; \sigma)}{p_v(\mathbf{o}|\mathbf{q}; \sigma)} - \lambda \frac{p_{\bar{\theta}}(\mathbf{o}|\mathbf{q}; \sigma)}{p_v(\mathbf{o}|\mathbf{q}; \sigma)} \right) [-\log p_\theta(\mathbf{o}|\mathbf{q})], \quad \forall \lambda \in \mathbb{R}.$$

Theoretically, there is an optimal choice of λ that minimizes the variance. The natural choice of $\lambda = 1$ means implicitly matching the probability $p_\theta(\mathbf{o}|\mathbf{q}; \sigma)$ to fit $p_*(\mathbf{o}|\mathbf{q}; \sigma)$, which corresponds to our model weight baseline (15). When the frequency for sampling buffer F is small, we can assume $p_\theta(\mathbf{o}|\mathbf{q}; \sigma)$ does not deviate too much from $p_v(\mathbf{o}|\mathbf{q}; \sigma)$, thus this ratio should be close to 1, which corresponds to our group weight baseline (13). Finally, as we actually use the negative ELBO $\mathcal{L}_\theta(\mathbf{o}|\mathbf{q})$ instead of $-\log p_\theta(\mathbf{o}|\mathbf{q})$ in computing the loss, the variance reduction only holds *approximately*.

B.4. Proofs for the Weighted Direct Discriminative Optimization Objective

For notational simplicity, we ignore the conditional dependence on \mathbf{q} . Write

$$\mathcal{F}(p_\theta) = -\mathbb{E}_{p_*} \log \frac{p_\theta}{p_\theta + p_v} - \mathbb{E}_{p_v} \log \frac{p_v}{p_\theta + p_v}.$$

For any fixed \mathbf{o} , consider the function

$$p_\theta(\mathbf{o}) \mapsto -p_*(\mathbf{o}) \log \frac{p_\theta(\mathbf{o})}{p_\theta(\mathbf{o}) + p_v(\mathbf{o})} - p_v(\mathbf{o}) \log \frac{p_v(\mathbf{o})}{p_\theta(\mathbf{o}) + p_v(\mathbf{o})}.$$

The derivative with respect to $p_\theta(\mathbf{o})$ is $-\frac{p_*(\mathbf{o})}{p_\theta(\mathbf{o})} + \frac{p_*(\mathbf{o}) + p_v(\mathbf{o})}{p_\theta(\mathbf{o}) + p_v(\mathbf{o})}$, which is > 0 if $p_\theta(\mathbf{o}) > p_*(\mathbf{o})$ and < 0 if $p_\theta(\mathbf{o}) < p_*(\mathbf{o})$. Therefore, this function is minimized at $p_\theta(\mathbf{o}) \leftarrow p_*(\mathbf{o})$, which completes the proof.

C. Details of Experiments and Further Results

C.1. Introduction of Datasets and Rewards used

To ensure a fair comparison, we use the same datasets and training rewards as d1 (Zhao et al., 2025a). For a self-contained presentation, we list the datasets and the rewards below.

GSM8K. GSM8k (Cobbe et al., 2021) is a mathematical reasoning dataset featuring multi-step grade school math problems. We conduct fine-tuning on the train split and evaluate on the test split.⁴

The reward is decomposed as follows:

1. *XML Structure Reward*: +0.125 for each correctly placed opening and closing tag (`<reasoning>`, `</reasoning>`, `<answer>`, `</answer>`) and -0.001 for each extra token after the closing tag `</answer>`.
2. *Soft Format Reward*: +0.5 for responses matching the pattern `<reasoning>...</reasoning><answer>...</answer>`.
3. *Strict Format Reward*: +0.5 for matching the specified format precisely with correct line breaks.
4. *Integer Answer Reward*: +0.5 if the retrieved answer parses as an integer.
5. *Correctness Reward*: +2 if the returned answer equals the ground truth exactly.

MATH500. MATH500 (Lightman et al., 2023) is a mathematical reasoning dataset, as well as a curated collection of 500 high-school-level problems sampled from the MATH (Hendrycks et al., 2021) dataset. We conduct fine-tuning on the train split and evaluate on the test split.⁵

The reward comprises

1. *Format Reward*: 1 when answer tags are present and `\boxed` appears inside them; 0.75 when the tags are present but `\boxed` is absent; 0.50 when the tags are missing but `\boxed` is present; 0.25 when neither the tags nor `\boxed` appear.
2. *Correctness Reward*: +2 when the correct answer is enclosed in `\boxed{}`.

Countdown. Countdown (Pan et al., 2025) is a planning task that requires solving a combinatorial arithmetic challenge, which is to form a target number using basic arithmetic operations with a provided set of 3 numbers, where each number can only be used once. We train on the training split of the dataset from the TinyZero project (Pan et al., 2025), restricting to instances that use only three numbers, and evaluate on 256 synthetically generated countdown questions with three numbers.

The reward checks if an arithmetic expression constructed from given numbers reaches a target value. More specifically, it is 1 when the equation equals the target and uses exactly the available numbers, 0.1 when the equation uses the right numbers but does not reach the target, and 0 if otherwise.

Sudoku. Sudoku is a planning task that requires solving 4×4 Sudoku puzzles, which demand constraint satisfaction and logical elimination to correctly fill the grid. We use the training dataset from <https://github.com/Black-Phoenix/4x4-Sudoku-Dataset>, in particular, the subset containing one million unique puzzles, which was synthetically generated using code from Arel (2025). For evaluation purposes, we randomly generate 256 Sudoku puzzles using this generator. The reward equals the fraction of originally blank cells that the model fills correctly.

C.2. Training Hyperparameters and Evaluation

We choose the training hyperparameters following Zhao et al. (2025a) for a fair comparison. We also use the Transformer Reinforcement Learning library (TRL, von Werra et al. (2020)) to implement DMPO. During training, we also employed the same Low-Rank Adaptation (LoRA, Hu et al. (2022)) with a rank of $r = 128$ and scaling factor $\alpha = 64$. For all tasks, the training was conducted on 8 NVIDIA H100 or H200 GPUs with the hyperparameters described below.

⁴<https://huggingface.co/datasets/openai/gsm8k>

⁵<https://huggingface.co/datasets/ankner/math-500>

We use a maximum generation length 256 tokens, a batch size of 8 per GPU, and gradient accumulation steps of 2, and 16 generated rollouts per prompt. We optimized the model using the AdamW optimizer (Loshchilov & Hutter, 2019) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay of 0.1, learning rate of 3×10^{-6} , and gradient clipping at 0.2. For each clean sequence, we sampled 4 partially masked tokens to compute the WDCE/WDDO loss. For rollouts generation during training, we use a semi-autoregressive random order sampler (with temperature 0) and fast-dLLM (with temperature 0.2) with a block size of 32 to generate diverse responses, which is the recommended practice for using LLaDA series models as is described in Nie et al. (2025b). We train 4,000 steps (number of gradient updates) for GSM8K and MATH500, Countdown, and Sudoku, respectively.

In Tab. 1, DMPO-LLaDA uses the group weight baseline (13) on GSM8K, MATH500, and Sudoku, and the individual weight baseline (14) on Countdown. DMPO-LLaDA-SFT and DMPO-LLaDA-1.5 adopt the individual weight baseline (14) in all cases.

For the reproduction of the d1 results, we follow the guidelines listed in Zhao et al. (2025a) and first perform SFT on s1k (Muennighoff et al., 2025) before applying diffu-GRPO. We use the recommended hyperparameter setups and train for up to 13,000 iterations on each dataset before evaluating the results.

For computational efficiency, we use Flash Attention 2 (Dao, 2024) and 4-bit quantization. All experiments on DMPO share these hyperparameters. The main result reported in Tab. 1 used the group weight baseline defined in (13). The ablation study in Fig. 4 also follows the same set of hyperparameters above, except for using different choices of weight baselines.

For the evaluation of LLaDA-series checkpoints, we consider three different generation lengths: 128, 256, and 512. We correspondingly use 128, 256, and 512 steps for generation. For the LLaDA series of models, such as LLaDA-Instruct, LLaDA-1.5, d1-LLaDA, and our own DMPO-LLaDA, we employ the semi-autoregressive sampler with a block size of 32, a greedy decoding scheme with a temperature of 0, and the top- k remasking scheme to achieve the best inference results. For the Dream model, we report generation lengths 256 and 512, and employ the recommended practice with temperature 0.95 and the top- k remasking scheme.

C.3. Further Experimental Results

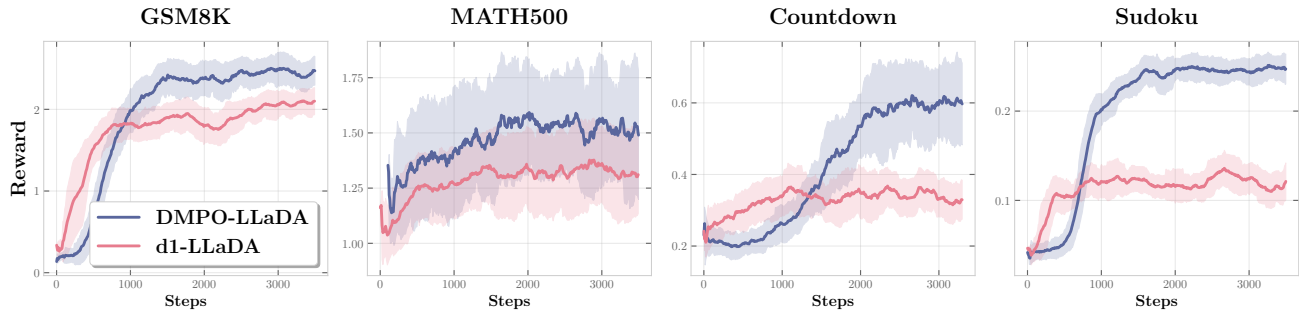


Figure 6. Reward dynamics during training. DMPO consistently produces higher rewards than d1.

DMPO consistently achieves higher rewards In Fig. 6, we present the reward dynamics of DMPO across training steps and compare with that of d1. DMPO consistently achieves higher reward values after an initial warm-up phase and ultimately discovers responses with higher rewards than d1, possibly because it continuously explores the reward distribution landscape throughout training. In the first 1,000 steps, DMPO often produces lower reward values than d1, potentially due to the lack of an SFT phase before RL scaling. Moreover, we observe that the performance of DMPO does not saturate after 4,000 gradient steps, suggesting its greater potential than GRPO-type algorithms.

Ablation studies on the hyperparameter dependence We provide an ablation study on two of the main hyperparameters in Alg. 1, namely the number of rollouts N and the frequency for sampling buffer F , in Figs. 7 and 8, respectively. For each run shown in Fig. 7, we train for 6 hours using 8 NVIDIA H200 GPUs. For each run shown in Fig. 8, we train for 8 hours with 8 NVIDIA H200 GPUs. We only vary the resampling buffer frequency F and the number of rollouts sampled per prompt N , while fixing other hyperparameters, such as the total effective batch size, to maintain a fair comparison.

For the number of rollouts per question N , we observe that a larger number of N does not necessarily lead to longer training

time, even with the same number of steps, due to the parallelism of the generation process, since we kept the total batch size fixed while varying the hyperparameter N . The algorithm is robust across various values of N ranging from 4 to 32 thanks to the mechanism for inserting negative gradients.

For the buffer sampling frequency F , we observe that it significantly affects training speed. The figure clearly demonstrates the advantage of DMPO due to its *off-policy* nature, whereas a purely on-policy realization of WDCE loss (with $F = 1$) is not only extremely slow but also does not show a significant boost in per-step reward gains. The figure also underscores the unique benefit of WDCE being a *forward* loss: given the generated rollouts and their weights, one can train using the simple forward process via random masking. Our algorithm is robust to choices of F up to 24, whereas an even larger F may cause slight instability later in training when the reward is high.

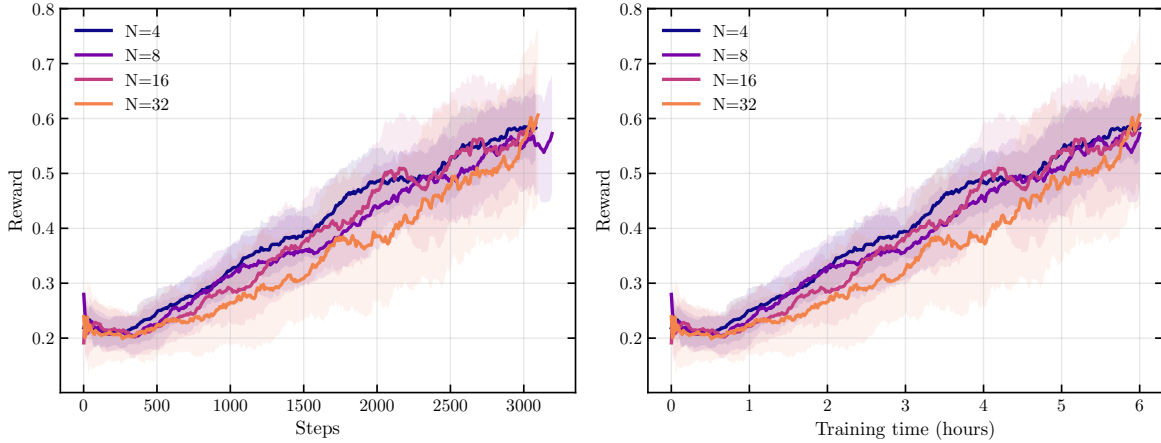


Figure 7. Ablation study of the number of rollouts per prompt N on Countdown dataset under the same training time and compute. The performance is robust to this hyperparameter.

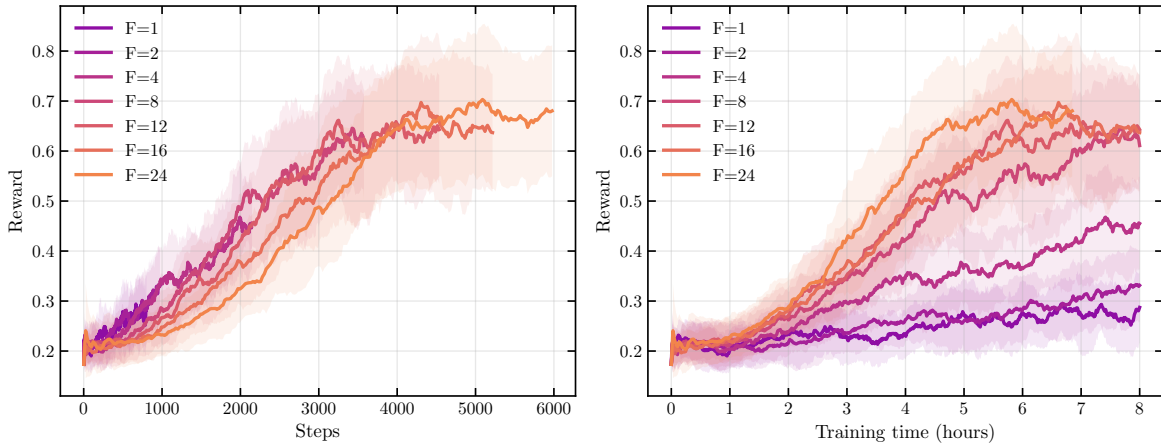


Figure 8. Ablation study of the resampling frequency F on Countdown dataset. A larger F is generally more time-efficient though may cause instability when the reward is high.

Visualizing rollout entropy of DMPO In Fig. 9, we compare the reward and entropy of the generated rollouts during training for both the relative-entropy-based (diffu-GRPO) RL algorithm and the cross-entropy-based (DMPO) RL algorithm. Here, in both experiments, we fix $N = 16$ and $F = 8$ and evaluate the entropy of generated samples every 10 generations. The evaluation of entropy is as follows: we use random-order autoregressive generation with block length 32, and at the d -th step of unmasking (where d ranges from 1 to $D = |\mathcal{o}|$), we compute the entropy of the predicted logits at the d -th position, and take average of all the D entropy values as the final value of sequential entropy. From the figure, the trend of consistently higher sample entropy for WDCE loss than for diffu-GRPO agrees with our expectation that cross-entropy-based methods are less prone to mode-seeking and maintain a higher level of diversity throughout training.

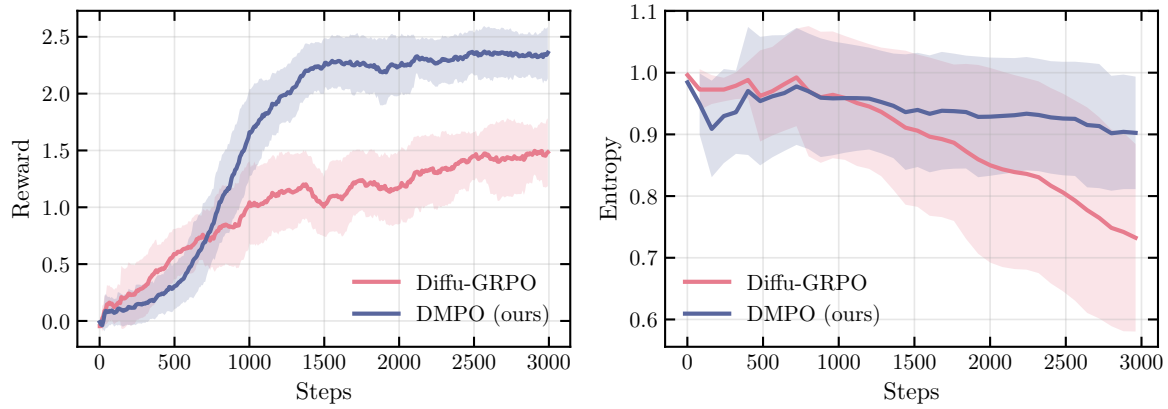


Figure 9. Reward (left) and entropy (right) of generated rollouts during training for both diffu-GRPO (Zhao et al., 2025a) and DMPO.

C.4. Example Outputs of the Model after Fine-tuning

We present two example outputs of the DMPO-LLaDA model in the following.

Model: DMPO-LLaDA

Dataset: MATH500

Generation Sequence Length: 512

Question: The product of $3t^2 + 5t + a$ and $4t^2 + bt - 2$ is $12t^4 + 26t^3 - 8t^2 - 16t + 6$. What is $a + b$?

To find the values of a and b such that the product of $3t^2 + 5t + a$ and $4t^2 + bt - 2$ equals $12t^4 + 26t^3 - 8t^2 - 16t + 6$, we need to expand the left-hand side and compare coefficients with the right-hand side.

First, expand the product:

$$(3t^2 + 5t + a)(4t^2 + bt - 2)$$

Using the distributive property (FOIL method for binomials):

$$3t^2 \cdot 4t^2 + 3t^2 \cdot bt + 3t^2 \cdot (-2) + 5t \cdot 4t^2 + 5t \cdot bt + 5t \cdot (-2) + a \cdot 4t^2 + a \cdot bt + a \cdot (-2)$$

Simplify each term:

$$12t^4 + 3bt^3 - 6t^2 + 20t^3 + 5bt^2 - 10t + 4at^2 + abt - 2a$$

Combine like terms:

$$12t^4 + (3b + 20)t^3 + (-6 + 5b + 4a)t^2 + (-10 + ab)t - 2a$$

We need this to equal:

$$12t^4 + 26t^3 - 8t^2 - 16t + 6$$

By comparing coefficients, we get the following system of equations:

$$3b + 20 = 26 \implies 3b = 6 \implies b = 2$$

$$-6 + 5(2) + 4a = -8 \implies 4 + 4a = -8 \implies 4a = -12 \implies a = -3$$

Thus, $a = -3$ and $b = 2$.

$$a + b = -3 + 2 = -1$$

</reasoning><answer>

Model: DMPO-LLaDA

Dataset: GSM8K

Generation Sequence Length: 512

Question: Katarina has 5 less cookies than Max has. Max has 12 more cookies than the Cookie Monster, and Summer has 23 more cookies than Max. If Katarina has 68 cookies, how many cookies do they have in total?

To determine the total number of cookies Katarina, Max, and Summer have, we need to find out how many cookies each person has and then sum them up.

1. We know that Katarina has 68 cookies.
2. Katarina has 5 less cookies than Max.
3. Max has 12 more cookies than the Cookie Monster.
4. Summer has 23 more cookies than Max.

Let's denote the number of cookies each person has as follows:

- K for Katarina's cookies
- M for Max's cookies
- C for the Cookie Monster's cookies
- S for Summer's cookies

From the problem, we have:

- $K = 68$
- $K = M - 5$
- $M = C + 12$
- $S = M + 23$

First, we substitute the known value of K into the equations:

- $68 = M - 5$
- $M = C + 12$
- $S = M + 23$

First, solve for M :

$$\begin{aligned} 68 &= M - 5 \\ M &= 68 + 5 \\ M &= 73 \end{aligned}$$

Next, solve for C :

$$\begin{aligned} 73 &= C + 12 \\ C &= 73 - 12 \\ C &= 61 \end{aligned}$$

Finally, solve for S :

$$\begin{aligned} 73 &= S - 23 \\ S &= 73 + 23 \\ S &= 96 \end{aligned}$$

Now, we have the number of cookies each person has:

- Katarina: 68 cookies
- Max: 73 cookies
- The Cookie Monster: 61 cookies
- Summer: 96 cookies

To find the total number of cookies, we sum these amounts:

$$68 + 73 + 61 + 96 = 298$$

Thus, the total number of cookies is 298.

`</reasoning><answer>`298`</answer>`