Enhancing Reasoning for Diffusion LLMs via Distribution Matching Policy Optimization (DMPO)

Yuchen Zhu*, Wei Guo*, Jaemoo Choi, Petr Molodyk, Bo Yuan, Molei Tao†, Yongxin Chen†

Georgia Institute of Technology
ArXiv 2510.08233

Takeaway Messages

- **TL;DR**: weighted denoising cross-entropy (WDCE) – treating i.i.d. samples from the current policy as importance weighted samples from the optimal path measure, and apply the standard denoising cross-entropy loss for MDM.
- DMPO is an RL method suitable for diffusion LLMs (dLLMs), with theoretical justification from stochastic optimal control and practical benefits of *off-policy* and *forward-type loss*.

Preliminaries on dLLMs

- **Mask diffusion model (MDM)**: input partially masked sequence $\boldsymbol{x} = (x_1, ..., x_D) \in \{1, 2, ..., V, \mathsf{M}\}^D$, output $\boldsymbol{\pi}_\theta(\boldsymbol{x}) \in \mathbb{R}^{D \times V}$, where each row is a probability vector over the vocabulary:

$$\boldsymbol{\pi}_\theta(\boldsymbol{x})_{d,u} \begin{cases} = 1_{x_d = u}, & x_d \neq \mathsf{M}, \\ \approx \Pr_{\mathrm{X} \sim p_{\text{data}}}(X_d = u | \boldsymbol{X}_{\text{UM}} = \boldsymbol{x}_{\text{UM}}), & x_d = \mathsf{M}. \end{cases}$$

- (Clean) sequence probability is defined by **random-order autoregressive (ROAR)** generation: $\boldsymbol{\sigma}$ is a random permutation of $\{1, ..., D\}$,

$$p_\theta(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\sigma}}\, p_\theta(\boldsymbol{x}; \boldsymbol{\sigma}) \text{ where } p_\theta(\boldsymbol{x}; \boldsymbol{\sigma}) = \prod_{d=1}^{D} \boldsymbol{\pi}_\theta(x_{\sigma_d} | \boldsymbol{x}_{\boldsymbol{\sigma}_{<d}}).$$

- **ELBO** approximation of sequence log-likelihood:

$$\begin{aligned} &-\log p_\theta(\boldsymbol{x}) = -\log \mathbb{E}_{\boldsymbol{\sigma}}\, p_\theta(\boldsymbol{x}; \boldsymbol{\sigma}) \leq -\mathbb{E}_{\boldsymbol{\sigma}} \log p_\theta(\boldsymbol{x}; \boldsymbol{\sigma}) \\ &= \mathop{\mathbb{E}}_{m \sim \mathrm{Unif}\{1, \ldots, |\boldsymbol{x}|\}} \left[\frac{|\boldsymbol{x}|}{m}\, \mathbb{E}_{\mu_m(\widetilde{\boldsymbol{x}} | \boldsymbol{x})} \sum_{d : \widetilde{x}_d = \mathsf{M}} -\log \boldsymbol{\pi}_\theta(\widetilde{\boldsymbol{x}})_{d, x_d} \right] =: \mathcal{L}_\theta(\boldsymbol{x}). \end{aligned}$$

RL for dLLMs: Problem Formulation

- **Aim**: Given a pretrained dLLM policy $\boldsymbol{\pi}_{\text{ref}}(\boldsymbol{o}|\boldsymbol{q})$ that samples from a data distribution $p_{\text{ref}}(\boldsymbol{o}|\boldsymbol{q})$, a reward function $r : (\boldsymbol{q}, \boldsymbol{o}) \mapsto \mathbb{R}$, a set of prompts $\mathcal{D}$, and temperature $\alpha > 0$, learn a dLLM policy $\boldsymbol{\pi}_\theta(\boldsymbol{o}|\boldsymbol{q})$ to produce the desired optimal sequence distribution

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

- **Key challenge**: mismatch between dLLM policy distribution and sample distribution, unlike AR LLMs.
- Not reward maximization but distribution matching – avoid mode collapse, maintain diversity.
- We also choose **cross-entropy loss** $\mathrm{KL}(p_* \| p_\theta)$ over relative-entropy $\mathrm{KL}(p_\theta \| p_*)$ to avoid mode covering.

Weighted Denoising Cross-Entropy (WDCE) Loss

- WDCE:

$$\min_{\theta} \mathop{\mathbb{E}}_{q\sim\mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \mathop{\mathbb{E}}_{p_{\text{old}}(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})} \left\{ w(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma}) \mathop{\mathbb{E}}_{m\sim\text{Unif}\{1,\dots,|\boldsymbol{o}|\}} \left[\frac{|\boldsymbol{o}|}{m} \mathop{\mathbb{E}}_{\mu_m(\widetilde{\boldsymbol{o}}|\boldsymbol{o})} \sum_{d:\widetilde{o}_d=\mathsf{M}} -\log \boldsymbol{\pi}_\theta(\widetilde{\boldsymbol{o}}|\boldsymbol{q})_{d,o_d} \right] \right\},$$

$$\text{where } w(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma}) = \frac{p_*(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})}{p_{\text{old}}(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})} \propto \exp\left(\frac{r(\boldsymbol{q},\boldsymbol{o})}{\alpha} + \log \frac{p_{\text{ref}}(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})}{p_{\text{old}}(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})} \right).$$

- **Core idea**: treat i.i.d. samples from the current policy as importance weighted samples from the optimal path measure, and apply the standard denoising cross-entropy loss for MDM.
- **Benefits**: *off-policy* (p_{old} can be any policy without gradient), *forward-type loss* (no need to keep track of sampling trajectory, only need final sample and their weights with simple masking).
- **Intuition from stochastic optimal control**: WDCE minimizes the KL divergence between the optimal and current path measures of the corresponding CTMCs.

Modifications to WDCE Loss for dLLMs

- The loss was initially proposed for **learning neural samplers** for a given target distribution $\pi \propto \mathrm{e}^{-U}$.[1]
- For dLLMs, needs some slight modifications:
- **Problem I**: All weights $w(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma}) = \frac{p_*(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})}{p_{\text{old}}(\boldsymbol{o}|\boldsymbol{q};\boldsymbol{\sigma})}$ are positive – promoting all responses even with low rewards. When batch size is small, this may penalize the likelihood of other unseen good responses to maintain a valid probability.
- **Modification I**: simply subtract a baseline (mean of all weights) to allow for negative gradients.
- **Problem II**: *Exact* computation of weights requires *exact* simulation of the CTMC (NFE = response length), too expensive.
- **Modification II**: We found that what really matters is the relative magnitudes of the weights. Use ELBO to approximate sequential log-likelihood after generation $\implies$ also allows us to use fast samplers such as with approximate KV caching, e.g., FastdLLM.[2]

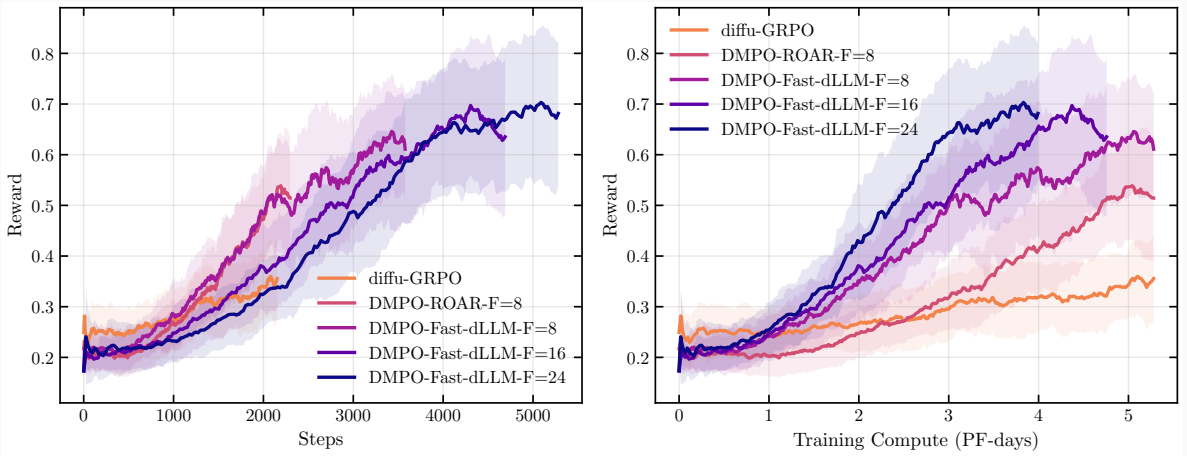
[1] Zhu et al. MDNS: Masked Diffusion Neural Sampler via Stochastic Optimal Control. NeurIPS 2025.

[2] Wu et al. Fast-dLLM: Training-free Acceleration of Diffusion LLM by Enabling KV Cache and Parallel Decoding. ICLR 2026.

A Glimpse at the Experimental Results

Table 1. Model performances on reasoning benchmarks. **Best** and second best results are highlighted. DMPO consistently outperforms other baselines across different generation length.

Task	GSM8K			MATH500			Countdown			Sudoku		
Sequence Length	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
DMPO-LLaDA (Ours)	74.83	82.41	**85.22**	30.00	38.20	**42.80**	**67.19**	**80.86**	82.81	**32.76**	24.56	19.97
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



See more details and ablations in the paper!