# Reinforcement Learning for Diffusion LLMs via Energy-Based Gibbs Alignment

**Yijia Fan[1], Jing Yang[1], Mingyu Liu[1], Kaitong Cai[1],**
**Jian Wang[2], Keze Wang[1]*, Jusheng Zhang[1]***
[1]Sun Yat-sen University, [2]Snap Inc.
{kezewang, jushengzhang88889}@gmail.com

## Abstract

Diffusion Large Language Models (dLLMs) have emerged as a promising non-autoregressive paradigm for text generation, offering parallel decoding and bidirectional context modeling. However, aligning dLLMs with reinforcement learning (RL) remains a significant challenge, as the marginal likelihood of sequences in masked diffusion is typically intractable, rendering standard policy gradient methods unstable or computationally prohibitive. In this work, we propose **Diffusion-Gibbs Alignment (DGA)**, a novel variational framework that reformulates RL for dLLMs as a distribution matching problem. DGA bypasses the explicit computation of log-probabilities by leveraging a learned energy function to model the relative quality of samples. The optimization is decoupled into two stable steps: (1) contrastive energy ranking to capture global reward structures, and (2) weighted diffusion alignment to update the policy via importance sampling. Empirically, DGA establishes a new state-of-the-art across logical reasoning (Sudoku, Countdown), mathematical reasoning (GSM8K, Math500), and code generation (HumanEval, MBPP) benchmarks. DGA offers a novel variational perspective for dLLM alignment, achieving better performance while simultaneously enhancing training speed and memory efficiency.

## 1 Introduction

Diffusion Large Language Models (dLLMs) (Nie et al., 2025; Li et al., 2025; Yu et al., 2025; Yang et al., 2025b) have recently emerged as a promising non-autoregressive paradigm for text generation, offering parallel decoding capabilities and flexible bidirectional context modeling unlike their Autoregressive (AR) counterparts (Radford et al., 2019; Touvron et al., 2023a; Yang et al., 2025a). Concurrently, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and direct RL fine-tuning strategies (Schulman et al., 2017; Rafailov et al., 2024; Zhang et al., 2024) have proven instrumental in unlocking the reasoning and alignment capabilities of AR models, as evidenced by the success of models like InstructGPT (Ouyang et al., 2022) and DeepSeek-R1 (Deepseek, 2025). Given this trajectory, a natural and compelling research direction is to extend these powerful reinforcement learning strategies to dLLMs, aiming to combine the structural advantages of diffusion generation with the goal-directed optimization of RL.

However, applying standard RL algorithms to dLLMs presents fundamental challenges. In AR models, the sequence log-probability $\log p_\theta(x|q)$ is trivially decomposable via the chain rule, enabling efficient gradient estimation for methods like PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024). In contrast, for Masked Diffusion Models such as LLaDA (Nie et al., 2025), the marginal likelihood of a sequence requires integrating over all possible permutation paths of the denoising process (Ho et al., 2020; Lipman et al., 2023), rendering the exact computation of $\log p_\theta(x|q)$ intractable. Consequently, the gradient of the log-probability $\nabla_\theta \log \pi_\theta$, a cornerstone of policy gradient methods, cannot be directly accessed. Existing approaches attempt to circumvent this by introducing heavy machinery such as tree-search rollouts (Pan et al., 2025) or relying on mean-field approximations (Jindal et al., 2025). These methods often incur significant computational overhead during training and inference. More critically, inaccurate probability estimation can lead to optimization instability, manifesting as weight collapse or a sudden drop in generation entropy, which severely hampers the model's exploration capability and diversity.

To address these limitations, we rethink the

* Corresponding authors.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45938–45948

dLLM policy optimization problem from an energy-based perspective (Du and Mordatch, 2020; Du et al., 2024; Xu et al., 2025). We observe that while the exact log-probability is inaccessible, the *relative quality* of samples can be effectively modeled without normalization constants. This insight allows us to bypass the explicit calculation of $\log p_\theta(x|q)$. Specifically, we propose to leverage a learned energy function to derive sample importance weights $w_i$. We show that weighting the standard masked diffusion loss with these energy-derived weights is mathematically equivalent to performing Empirical Risk Minimization (Brownlees and Guđmundsson, 2023; Lecué and Mendelson, 2016; Block et al., 2024) on a reward-reweighted data distribution. This formulation naturally aligns the diffusion model with high-reward regions without requiring the unstable estimation of policy gradients.

Based on this insight, we propose **Diffusion-Gibbs Alignment (DGA)**, which casts RL for dLLMs as variational distribution matching. DGA augments a diffusion policy with a prompt-conditioned residual energy $E_\phi(x, q)$, defining a Gibbs-optimal target $\pi^*(x|q) \propto p_{\text{prior}}(x|q) \exp(R(x, q)/\tau)$ and a joint policy $\pi_{\text{joint}}(x|q) \propto p_\theta(x|q) \exp(-E_\phi(x, q))$. Minimizing $D_{\text{KL}}(\pi_{\text{joint}} \| \pi^*)$ yields a tractable two-stage procedure: (i) learn $E_\phi$ via contrastive ranking within prompt groups, and (ii) update $\theta$ with *Weighted Diffusion Alignment* using reweighted samples under the standard diffusion training loop, avoiding tree search and log-probability evaluation. Experiments on challenging reasoning benchmarks show that DGA outperforms prior dLLM-RL baselines at the same rollout budget, with improved stability (higher ESS, lower seed variance, and controlled entropy) and lower inference cost by removing test-time tree search. Overall, DGA offers a rigorous and practical route to reliable non-autoregressive reasoning agents.

## 2 Related Work

**Diffusion Large Language Models.** The landscape of text generation has been long dominated by Autoregressive (AR) Large Language Models (LLMs) (Touvron et al., 2023a,b; llama team, 2024; Jiang et al., 2023), which generate tokens sequentially. While successful, AR models suffer from high inference latency and a lack of bidirectional context. Recently, Diffusion Large Language Models (dLLMs), particularly Masked Diffusion Models (MDMs) such as LLaDA (Nie et al., 2025), have emerged as a powerful non-autoregressive alternative. By modeling text generation as a denoising process, dLLMs enable parallel decoding and flexible bidirectional editing. However, unlike AR models where the likelihood is exactly decomposable via the chain rule, the marginal likelihood in dLLMs involves marginalizing over all possible denoising paths, making exact probability computation and subsequent optimization non-trivial.

**Reinforcement Learning for LLM Alignment.** Aligning LLMs with human preferences via Reinforcement Learning (RL) has become a standard paradigm, exemplified by Reinforcement Learning from Human Feedback (Ouyang et al., 2022). Standard algorithms like PPO (Schulman et al., 2017) and DPO (Rafailov et al., 2024) rely on the efficient calculation of log-probabilities or their gradients. For dLLMs, several recent works have attempted to bridge this gap. Approaches like d-TreeRPO (Pan et al., 2025) employ tree-search rollouts to estimate structural advantages, while others like Diffu-GRPO (Zhao et al., 2025) and VRPO (Zhu et al., 2025) utilize mean-field approximations or proxy probabilities. Despite their progress, these methods often incur heavy computational overhead or suffer from training instability (e.g., mode collapse) due to inaccurate gradient estimation in the diffusion space.

**Energy-Based Guidance in Diffusion.** Diffusion models are inherently connected to score-matching and energy-based models (EBMs) (Song and Kingma, 2021; Blondel et al., 2025). Guiding diffusion processes via external energy functions or classifiers has been widely explored in the image domain for controllable generation (Hong, 2024; Zhao et al., 2022). In the context of alignment, the control-as-inference framework (Levine, 2018) views the optimal policy as a Gibbs distribution reweighted by a reward-derived energy function. Recent works have explored using GFlowNets (Bengio et al., 2021) or reward-weighted regression to bypass explicit policy gradients. Our work, DGA, builds on this intuition by introducing a learned energy function $E_\phi(x, q)$ as a residual corrector to the dLLM. By reformulating the RL problem as a variational distribution matching task, we eliminate the need for intractable log-likelihood gradients and provide a stable, efficiency-focused alignment framework.

## 3 Method

We introduce Diffusion-Gibbs Alignment (DGA), a principled framework for aligning diffusion large language models (dLLMs) with differentiable or non-differentiable reward functions. We reformulate policy optimization as a variational distribution matching problem: a parameterized joint distribution is trained to approximate an optimal Gibbs distribution induced by task rewards. We tackle the resulting intractable matching via coordinate ascent, decoupling the optimization into (i) preference learning through contrastive energy ranking and (ii) policy optimization through weighted diffusion alignment.

### 3.1 Variational Alignment via Gibbs Distributions

Consider a prompt distribution $q \sim \mathcal{Q}$ and a diffusion policy $p_\theta(x|q)$ that generates a complete response trajectory $x \in \mathcal{X}$. Our goal is to align $p_\theta$ so as to maximize a reward signal $R(x, q)$. Following the control-as-inference view, we define the optimal target policy $\pi^*$ as a Gibbs distribution, which reweights a prior distribution (or the initial policy) $p_{\text{prior}}(x|q)$ by the exponentiated reward:

$$\pi^*(x|q) = \frac{1}{Z^*(q)} p_{\text{prior}}(x|q) \exp\left(\frac{R(x,q)}{\tau}\right), \tag{1}$$

where $\tau > 0$ is a temperature hyperparameter and $Z^*(q)$ is the partition function for the optimal policy.

To approximate this target, we introduce a parameterized joint policy $\pi_{\text{joint}}$ defined by augmenting the current diffusion model $p_\theta$ with a learnable energy function $E_\phi(x, q)$. The energy acts as a residual correction to the base probability:

$$\pi_{\text{joint}}(x|q;\theta,\phi) = \frac{1}{Z_{\text{joint}}(q)} p_\theta(x|q) \exp\left(-E_\phi(x,q)\right), \tag{2}$$

where $Z_{\text{joint}}(q) = \int p_\theta(x|q) \exp(-E_\phi(x,q))dx$ is the (intractable) partition function of the joint policy.

We aim to minimize the Kullback-Leibler (KL) divergence between the parameterized joint policy and the optimal target policy:

$$\min_{\theta,\phi} \mathcal{L}_{\text{KL}}(\theta,\phi) = \mathbb{E}_{q\sim\mathcal{Q}}\left[D_{\text{KL}}(\pi_{\text{joint}}(\cdot|q) \,\|\, \pi^*(\cdot|q))\right]. \tag{3}$$

Directly minimizing Eq. (3) is challenging due to the normalizing constants. However, we can transform this reverse-KL minimization into an equivalent variational maximization objective.

**Proposition 1.** *Minimizing the reverse KL divergence in Eq. (3) is equivalent to maximizing the following variational objective $\mathcal{J}(\theta,\phi)$, up to a constant independent of $\theta$ and $\phi$:*

$$\begin{aligned}\mathcal{J}(\theta,\phi) &= \mathbb{E}_{q\sim\mathcal{Q}}\left[\mathbb{E}_{x\sim\pi_{\text{joint}}}\left[\log\frac{\pi^*(x\mid q)}{\pi_{\text{joint}}(x\mid q)}\right]\right] \\ &= \mathbb{E}_{q\sim\mathcal{Q}}\left[\mathbb{E}_{x\sim\pi_{\text{joint}}}\left[\frac{R(x,q)}{\tau} + \log\frac{p_{\text{prior}}(x\mid q)}{p_\theta(x\mid q)} + E_\phi(x,q)\right] - \log Z_{\text{joint}}(q)\right] \\ &= \mathbb{E}_{q\sim\mathcal{Q}}\left[\mathbb{E}_{x\sim\pi_{\text{joint}}}\left[\frac{R(x,q)}{\tau} + \log\frac{p_{\text{prior}}(x\mid q)}{p_\theta(x\mid q)} + E_\phi(x,q)\right] - \log Z_{\text{joint}}(q)\right].\end{aligned} \tag{4}$$

*Proof Sketch.* Expanding the KL definition, we have $D_{\text{KL}}(\pi_{\text{joint}}\|\pi^*) = \mathbb{E}_{\pi_{\text{joint}}}[\log\pi_{\text{joint}} - \log\pi^*]$. Substituting Eq. (1) and Eq. (2), we obtain:

$$\begin{aligned}\log\pi_{\text{joint}} - \log\pi^* = \left(\log p_\theta - E_\phi - \log Z_{\text{joint}}\right) \\ - \left(\log p_{\text{prior}} + \frac{R}{\tau} - \log Z^*\right).\end{aligned} \tag{5}$$

Since $\log Z^*(q)$ is constant with respect to $(\theta,\phi)$, minimizing the expectation of the above expression is equivalent to maximizing $\mathcal{J}(\theta,\phi)$. Crucially, $\log Z_{\text{joint}}(q)$ depends on both $\theta$ and $\phi$, so optimization is non-trivial. We therefore adopt a Coordinate Ascent strategy, alternating between updating $\phi$ (Preference Learning) and $\theta$ (Policy Optimization).

### 3.2 Preference Learning via Contrastive Energy Ranking

In the first phase of coordinate ascent, we fix the policy parameters $\theta$ and update the energy parameters $\phi$. A common pitfall is to differentiate only the explicit terms in Eq. (4) while treating the sampling distribution $\pi_{\text{joint}}$ as fixed. Doing so yields the seemingly "self-canceling" expression because

$$\begin{aligned}\nabla_\phi \log Z_{\text{joint}}(q) &= \nabla_\phi \log \int p_\theta(x\mid q)\exp\big(-E_\phi(x,q)\big)\,dx \\ &= -\mathbb{E}_{x\sim\pi_{\text{joint}}}[\nabla_\phi E_\phi(x,q)].\end{aligned} \tag{6}$$

However, this cancellation is *not* a correct gradient of $\mathcal{J}$, because $\pi_{\text{joint}}(\cdot|q;\theta,\phi)$ itself depends on $\phi$ through Eq. (2). Formally, differentiating an expectation under a $\phi$-dependent distribution must include a score-function term:

$$\begin{aligned}\nabla_\phi\mathbb{E}_{x\sim\pi_{\text{joint}}}[f(x,\phi)] &= \mathbb{E}_{x\sim\pi_{\text{joint}}}[\nabla_\phi f(x,\phi)] \\ &\quad + \mathbb{E}_{x\sim\pi_{\text{joint}}}\big[f(x,\phi)\nabla_\phi\log\pi_{\text{joint}}(x\mid q)\big].\end{aligned} \tag{7}$$

Applying Eq. (7) to Eq. (4) shows that the full gradient $\nabla_\phi\mathcal{J}$ contains additional terms involving $\nabla_\phi\log\pi_{\text{joint}}$ and therefore does *not* trivially vanish.

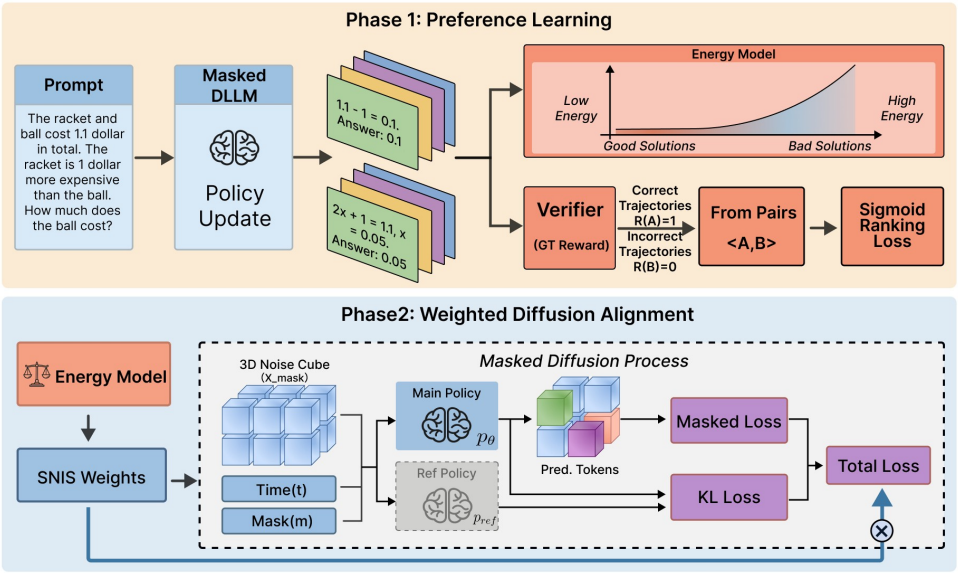


Figure 1: **Overview of the Diffusion-Gibbs Alignment (DGA) framework.** Phase 1 (top) performs preference learning via contrastive energy ranking to approximate the reward landscape. Phase 2 (bottom) executes policy optimization by aligning the masked diffusion process with importance-weighted gradients derived from the learned energy model.

In practice, directly optimizing $\phi$ via Eq. (4) is still difficult because it requires (i) sampling from $\pi_{\text{joint}}$ and (ii) handling the partition function effects inside $\log \pi_{\text{joint}}$, both of which are intractable for large-scale dLLMs.

Instead, we interpret $E_\phi$ as a discriminator / preference model and circumvent explicit partition-function estimation by learning from *pairwise preferences*. For a given prompt $q$, we sample a set of trajectories $\{x_i\}_{i=1}^K$ from the current policy and obtain their rewards $R_i$. We then construct preference pairs $\mathcal{P}(q) = \{(i,j) \mid R_i > R_j\}$. Assuming the preference probability follows a Bradley–Terry model parameterized by energy differences:

$$P(x_i \succ x_j|q) = \sigma\left(\frac{E_\phi(x_j,q) - E_\phi(x_i,q)}{\beta}\right), \tag{8}$$

where $\sigma$ is the sigmoid function and $\beta$ is a scaling factor, we optimize $\phi$ by minimizing the negative log-likelihood of the preference data:

$$\mathcal{L}_{\text{Energy}}(\phi) = -\mathbb{E}_{q\sim\mathcal{Q}}\mathbb{E}_{(i,j)\in\mathcal{P}(q)}\left[\log\sigma\left(\frac{E_\phi(x_j,q) - E_\phi(x_i,q)}{\beta}\right)\right]. \tag{9}$$

Minimizing Eq. (9) aligns the energy landscape with the reward structure such that lower energy corresponds to higher reward, i.e., $E_\phi(x,q) \approx -R(x,q)$. This ranking formulation is shift-invariant, meaning it avoids explicit dependence on the partition function $Z_{\text{joint}}(q)$ by considering only relative energy differences. Furthermore, this objective can be viewed as extracting a verifiable reward signal into a differentiable scalar, enabling stable downstream policy optimization.

### 3.3 Policy Optimization via Weighted Diffusion Alignment

In the second phase, we fix the energy model $\phi$ and update the diffusion policy $\theta$ to maximize $\mathcal{J}(\theta,\phi)$. Retaining only the terms dependent on $\theta$, the objective simplifies to:

$$\max_\theta \mathcal{J}_\theta = \mathbb{E}_{q\sim\mathcal{Q}}\Big[\mathbb{E}_{x\sim\pi_{\text{joint}}}[-\log p_\theta(x|q)] - \log Z_{\text{joint}}(q)\Big]. \tag{10}$$

Since sampling exactly from $\pi_{\text{joint}}$ is intractable, we employ Self-Normalized Importance Sampling (SNIS) using the current policy $p_{\theta_{\text{old}}}$ as the proposal distribution. The importance weights are derived

from the energy function:

$$w(x|q) = \exp\left(-\frac{E_\phi(x,q)}{\alpha}\right), \quad w_i = \frac{w(x_i|q)}{\sum_{j=1}^K w(x_j|q)} \tag{11}$$

where $\{x_i\}_{i=1}^K \sim p_{\theta_{\text{old}}}(\cdot|q)$ are sampled candidates. The gradient of Eq. (10) can be approximated by reweighting the gradient of the log-likelihood:

$$\nabla_\theta \mathcal{J}_\theta \approx \mathbb{E}_q \sum_{i=1}^K w_i \nabla_\theta \log p_\theta(x_i|q). \tag{12}$$

However, for Masked Diffusion Models (MDMs) such as LLaDA, the exact log-probability $\log p_\theta(x|q)$ involves integrating over all decoding permutations $\sigma \in S_L$, which is computationally intractable:

$$p_\theta(x|q) = \mathbb{E}_{\sigma \sim S_L}\left[\prod_{t=1}^T p_\theta(x_{\sigma_t}|x_{\sigma_{<t}}, q)\right]. \tag{13}$$

To resolve this, we propose *Weighted Diffusion Alignment*. We observe that the standard diffusion training loss $\mathcal{L}_{\text{MDM}}$ (typically masked cross-entropy) provides a practical surrogate for optimizing the intractable marginal likelihood. In particular, we optimize a weighted diffusion objective:

$$\min_\theta \mathcal{L}_{\text{Align}}(\theta) = \mathbb{E}_q \sum_{i=1}^K w_i \cdot \mathbb{E}_{t,m}\left[\mathcal{L}_{\text{MDM}}(x_i, q, t, m; \theta)\right] \tag{14}$$

where $\mathcal{L}_{\text{MDM}}$ is the loss computed on masked tokens $m$ at timestep $t$ (with $t \sim U(0,1)$ and $m \sim p_{\text{mask}}(t)$). By optimizing Eq. (14), we push the probability mass of $p_\theta$ toward regions with lower energy (higher reward) without explicitly computing the intractable marginal $\log p_\theta(x|q)$.

### 3.4 Practical Implementation: Trust Regions and Algorithm

To ensure training stability and prevent policy collapse, we introduce a local trust region constraint. Unlike autoregressive models where KL divergence is computed over the entire sequence, we impose a masked KL penalty computed only at the denoising positions of the current timestep $t$. The regularized objective is:

$$\mathcal{L}_{\text{Total}}(\theta) = \sum_{i=1}^K w_i \Big( \mathcal{L}_{\text{MDM}}(x_i, \dots) + \lambda \mathcal{D}_{\text{KL}}^{\text{masked}}(p_\theta(\cdot|x_t^{\text{mask}}) \| p_{\text{ref}}(\cdot|x_t^{\text{mask}})) \Big). \tag{15}$$

**Algorithm 1** Diffusion-Gibbs Alignment (DGA)

**Input:** Prompt distribution $\mathcal{Q}$, Base Policy $\theta$, Energy Model $\phi$, Reference $\theta_{\text{ref}}$.
**Hyperparameters:** Rollout $K$, Temp $\alpha, \beta$, Trust $\lambda$.
**for** iteration $n = 1$ **to** $N$ **do**
  **// Step 1: Exploration & Ranking**
  Sample prompts $q \sim \mathcal{Q}$.
  Generate trajectories $\{x_i\}_{i=1}^K \sim p_{\theta_{\text{old}}}(\cdot|q)$.
  Compute rewards $R_i = \text{Verifier}(x_i, q)$.
  Update $\phi$ via pairwise ranking loss $\mathcal{L}_{\text{Energy}}$ (Eq. 9).
  **// Step 2: Weighting & Alignment**
  Compute energies $e_i = E_\phi(x_i, q)$.
  Compute SNIS weights $w_i \propto \exp(-e_i/\alpha)$.
  Clip weights and normalize to obtain $\tilde{w}_i$ (Eq. 16).
  Sample indices $k \sim \text{Categorical}(\tilde{w}_1, \dots, \tilde{w}_K)$.
  Sample time $t \sim U(0,1)$ and mask $m$.
  Update $\theta$ by minimizing $\mathcal{L}_{\text{Total}}$ (Eq. 15).
  Update reference policy $\theta_{\text{old}} \leftarrow \theta$ periodically.
**end for**

where $\mathcal{D}_{\text{KL}}^{\text{masked}}$ measures the divergence between the current policy and the reference model on the predicted distributions for masked tokens.

Additionally, to manage the variance of the importance weights $w_i$, we apply weight clipping with a factor $\gamma$:

$$\hat{w}_i = \min\left(w_i, \gamma \cdot \frac{1}{K}\right), \quad \tilde{w}_i = \frac{\hat{w}_i}{\sum_j \hat{w}_j}. \tag{16}$$

We monitor the Effective Sample Size $\text{ESS} = 1/\sum \tilde{w}_i^2$ to dynamically adjust the temperature $\alpha$ if the distribution becomes too peaked. The complete training procedure is summarized in Algorithm 1.

## 4 Experiments

In this section, we empirically validate the effectiveness of **Diffusion-Gibbs Alignment (DGA)** by comparing it against state-of-the-art reinforcement learning algorithms for diffusion large language models (dLLMs). We aim to demonstrate that DGA achieves superior performance across diverse reasoning and generation tasks without relying on computationally expensive tree-search structures.

### 4.1 Experimental Setup

**Models and Baselines.** To ensure a rigorous and fair comparison, we align our experimental settings with the current state-of-the-art method, d-TreeRPO (Pan et al., 2025). We utilize `LLaDA-8B-Instruct` (Nie et al., 2025) as the base model for all experiments. We compare DGA against a comprehensive suite of dLLM RL baselines, including **Diffu-GRPO** (Zhao et al., 2025), **VRPO** (Zhu et al., 2025), **wd1** (Tang et al., 2025), **SAPO** (Xie et al., 2025), and **d-TreeRPO** (Pan et al., 2025). Consistent with prior works, we employ LoRA (Hu et al., 2021) fine-tuning with rank $r = 128$ and alpha $\alpha = 64$. For DGA, the energy model $\phi$ shares the same architecture as the policy network but with a distinct projection head.

**Benchmarks and Metrics.** We evaluate performance on six challenging benchmarks categorized into three domains: (1) **Logical Reasoning**: *Sudoku* (4×4) and *Countdown*, following the settings in d-TreeRPO (Pan et al., 2025) to test strict constraint satisfaction; (2) **Mathematical Reasoning**: *GSM8K* (Cobbe et al., 2021) and *Math500* (Lightman et al., 2023), assessing multi-step reasoning capabilities; (3) **Code Generation**: To further evaluate the generalization capability of DGA beyond standard reasoning tasks, we introduce *HumanEval* (Chen et al., 2021) and *MBPP* (Austin et al., 2021). This represents a more challenging setting for dLLMs than previous works. All models are evaluated using **Pass@1** accuracy. We report results under two generation lengths: 256 and 512 tokens (where applicable), with 128 denoising steps.

### 4.2 Main Results

Table 1 presents the performance comparison of DGA with existing dLLM RL methods. **DGA establishes a new state-of-the-art across all evaluated benchmarks.** Specifically, on the constrained logical tasks (Sudoku and Countdown), DGA outperforms the previous best method, d-TreeRPO, by margins of **1.2%** and **2.4%** respectively, achieving near-perfect solution rates. This indicates that the energy-based guidance effectively captures the discrete constraints of the reward landscape without needing explicit tree-structured rollouts.

More importantly, DGA demonstrates significant superiority on complex reasoning and generation tasks. On **Math500**, which requires rigorous chain-of-thought reasoning, DGA achieves a score of **40.2%**, surpassing d-TreeRPO by **2.5%** and the base model by **+7.8%**. Furthermore, in the newly introduced code generation tasks, DGA exhibits remarkable generalization. On **HumanEval** and **MBPP**, DGA achieves **41.2%** and **47.8%** respectively, significantly outperforming the strong baseline d-TreeRPO by margins of **2.7%** and **3.6%**. These results suggest that by modeling the policy optimization as a variational Gibbs alignment problem, DGA avoids the approximation errors inherent in single-step probability estimation and the high variance of tree search, leading to a more robust and generalized policy even in code domains where structural correctness is paramount.

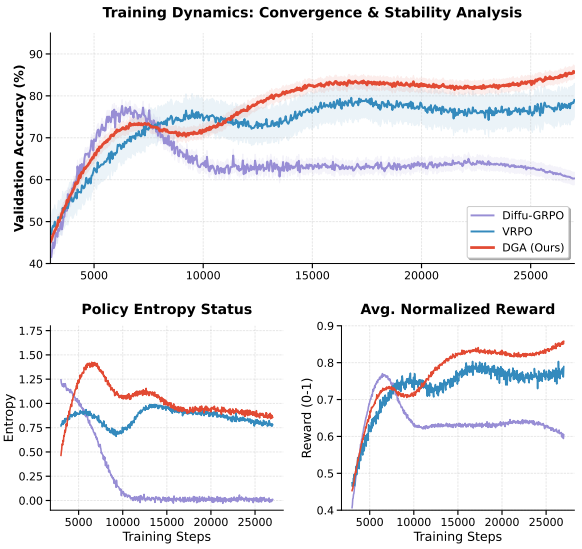


Figure 2: **Training Dynamics on GSM8K (0-27k Steps). Diffu-GRPO (Purple)** suffers from mode collapse (entropy $\to$ 0), leading to performance degradation. **VRPO (Blue)** exhibits persistent high-variance oscillations. **DGA (Red)** maintains healthy entropy levels and demonstrates robust convergence to the highest accuracy.

### 4.3 Training Stability & Convergence Analysis

To assess optimization stability and long-horizon convergence, we run DGA on *GSM8K* for 27,000 training steps. Since dLLM-RL is often unstable (e.g., mode collapse or high-variance oscillations), we compare DGA with **Diffu-GRPO** and the variance-reduced **VRPO**, tracking validation accuracy, policy entropy, and normalized reward over training.

As illustrated in Figure 2, DGA demonstrates superior stability compared to baselines. **Diffu-GRPO** (purple) exhibits a characteristic failure mode: it peaks early but suffers from *mode col-*

Table 1: Performance comparison of **DGA (Ours)** with existing dLLM RL methods across logical, mathematical, and coding benchmarks. All methods use `LLaDA-8B-Instruct` as the base model. Results are reported as **Pass@1 (%)**. The generation length is set to 256 for all tasks, with an additional 512 setting for math tasks. Top performance is **bolded**. Gray-shaded rows indicate the proposed method.

| Method | Sudoku | Countdown | GSM8K | | Math500 | | HumanEval | MBPP |
|---|---|---|---|---|---|---|---|---|
| | 256 | 256 | 256 | 512 | 256 | 512 | 256 | 256 |
| LLaDA-8B-Instruct | 6.7 | 19.5 | 76.7 | 78.2 | 32.4 | 36.2 | 35.4 | 40.0 |
| + Diffu-GRPO | 12.9 | 31.3 | 79.8 | 81.9 | 34.1 | 39.0 | 36.9 | 41.5 |
| + VRPO | 12.8 | 22.3 | 80.1 | 81.5 | 35.6 | 34.8 | 36.5 | 41.2 |
| + wd1 | 25.2 | 51.2 | 80.8 | 82.3 | 34.4 | 39.0 | 37.1 | 41.8 |
| + SAPO | 20.3 | 52.0 | 80.6 | 82.1 | 33.8 | 38.4 | 37.4 | 42.1 |
| + d-TreeRPO | 92.9 | 71.1 | 81.2 | 82.6 | 37.7 | 38.9 | – | – |
| **+ DGA (Ours)** | **94.1** | **73.5** | **83.5** | **84.8** | **40.2** | **42.1** | **41.2** | **47.8** |
| *Improvement* | (+1.2) | (+2.4) | (+2.3) | (+2.2) | (+2.5) | (+3.2) | (+3.8) | (+5.7) |

*lapse*, where policy entropy precipitously drops to zero (bottom-left), causing a severe decline in final performance. **VRPO** (blue) prevents total collapse but struggles with stability, showing jagged *oscillations* throughout training and failing to converge to an optimal policy. In contrast, **DGA** (red) maintains a healthy level of entropy, facilitating sustained exploration. Notably, DGA successfully navigates the optimization landscape—evident from the strategic adjustment phase around step 8k—and steadily converges to the state-of-the-art accuracy ($> 85\%$), confirming that our energy-based objective effectively balances exploration and exploitation.

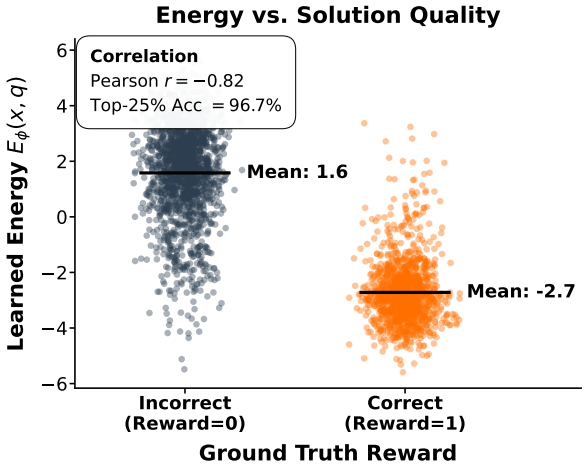


Figure 3: **Energy Score Distribution vs. Ground Truth Reward.** The plot reveals a distinct separation between incorrect samples (gray, Mean Energy $\approx 1.6$) and correct samples (orange, Mean Energy $\approx -2.7$).

## 5 In-depth Analysis

### 5.1 Analysis of Learned Energy Landscape

To validate the semantic meaningfulness of the learned energy function $E_\phi(x, q)$, we investigate whether the continuous energy landscape effectively captures the underlying discrete reward structure. In standard dLLM RL, the ground-truth reward $R(x, q)$ is typically sparse and non-differentiable (e.g., binary outcomes from a verifier), which often leads to high-variance gradient estimates. The core premise of DGA is that the energy function should act as a smooth, dense proxy for solution quality, where lower energy corresponds to higher correctness. To verify this, we randomly sampled $N = 2,000$ generated trajectories from the DGA policy trained on *GSM8K*. We visualized the distribution of their predicted energy scores against their ground-truth correctness labels to assess the alignment between the learned manifold and the task objective.

As shown in Figure 3, the energy model exhibits strong semantic alignment. Incorrect solutions (gray) concentrate at high energy ($\mu \approx 1.6$), while correct ones (orange) fall into a low-energy basin ($\mu \approx -2.7$), yielding a clear bimodal separation. Energy also correlates tightly with correctness (Pearson $r = -0.82$), and the lowest-energy quartile achieves **96.7%** accuracy. This explains the stability in Section 4.3: a smooth energy landscape provides dense learning signals under sparse rewards, steering the policy toward the correctness mode.

### 5.2 Ablation Study

To rigorously verify the contribution of each component in the proposed DGA framework, we performed a "leave-one-out" ablation study on the *GSM8K* dataset. We evaluated three variants by

Table 2: **Ablation Study on GSM8K.** We report Pass@1 Accuracy, the standard deviation of rewards ($\sigma_R$) to measure stability, and the Mean Effective Sample Size (ESS). The **Full DGA** row aligns with our main results. Removing the Energy Model causes the most significant drop (-5.6%), confirming that the continuous energy landscape is the primary driver of performance.

| Method Variant | GSM8K Acc. | $\Delta$ Acc. | Reward Std ($\sigma_R$) | Mean ESS | Failure Mode |
|---|---|---|---|---|---|
| **Full DGA (Ours)** | **84.8** | - | **0.85** | **94.2** | *None* |
| w/o Energy Model | 79.2 | -5.6 | 3.41 | 18.5 | *Sparse Gradient Signal* |
| w/o Masked KL | 81.5 | -3.3 | 2.95 | 88.1 | *Early Mode Collapse* |
| w/o Weight Clipping | 83.1 | -1.7 | 1.88 | 32.4 | *High Variance Updates* |

systematically removing key modules from the full implementation: (1) **w/o Energy Model**: Replacing the learned energy function $E_\phi(x, q)$ with the raw, normalized ground-truth reward for computing importance weights; (2) **w/o Masked KL**: Removing the local trust region constraint ($\lambda = 0$ in Eq. 15); (3) **w/o Weight Clipping**: Excluding the weight clipping mechanism (Eq. 16) to test numerical stability.

Table 2 summarizes the performance in terms of Pass@1 accuracy, reward stability ($\sigma_R$), and effective sample size (ESS).

**Analysis of Components.** The ablation results underscore the critical synergy among DGA's components. The learned energy landscape $E_\phi$ serves as the cornerstone; replacing it with raw rewards causes the most severe performance degradation (-5.6%) and a precipitous drop in Mean ESS to 18.5, confirming that the energy function acts as a necessary *dense smoothing operator* over sparse discrete rewards. Furthermore, the Masked KL constraint proves vital for optimization stability; its removal leads to a 3.3% accuracy drop and increased reward volatility ($\sigma_R = 2.95$), indicating that the trust region effectively prevents early mode collapse. Finally, weight clipping ensures numerical robustness (-1.7% impact); without it, gradient estimates are skewed by outliers (ESS drops to 32.4), whereas clipping maintains low-variance updates crucial for fine-grained convergence.

### 5.3 Computational Efficiency Analysis

Beyond generation quality, RL overhead is a key bottleneck for scaling dLLMs. Tree-search methods such as d-TreeRPO build rollout trees with branching factor $B$ and height $H$ to estimate advantages, leading to exponential compute and high memory usage. In contrast, DGA casts alignment as variational distribution matching and trains on independent sample batches without maintaining tree states. We quantify this on *GSM8K* with `8×A100`, reporting training hours, throughput (tokens/s), peak GPU memory, and the inference protocol used for the final results.

Table 3: **Computational Efficiency.** All methods use `LLaDA-8B` (20k steps / to convergence). **Rel. Cost** is normalized to Diffu-GRPO. DGA achieves SOTA with direct sampling; tree-based methods add training and/or test-time search overhead.

| Method | Time (Hours) | Rel. Cost | Memory (GB) | Throughput | Inference Protocol |
|---|---|---|---|---|---|
| Diffu-GRPO | 14.2 | 1.0× | 42.5 | 3,250 tok/s | Direct Sampling |
| d-TreeRPO | 52.8 | 3.7× | 78.2 | 890 tok/s | Tree Rollout |
| **DGA (Ours)** | **17.5** | **1.2×** | **46.8** | **2,640 tok/s** | **Direct Sampling** |

**Efficiency–Performance Trade-off.** As summarized in Table 3, DGA offers a stronger balance between efficiency and accuracy. Compared to the tree-based baseline d-TreeRPO, DGA is **3**× faster to train (17.5h vs. 52.8h) and uses **40%** less memory, since d-TreeRPO must store and backtrack gradients over the full rollout tree, which bottlenecks throughput (890 tok/s). Relative to Diffu-GRPO, DGA introduces only a modest **1.2**× overhead for energy estimation, yet delivers a clear gain (+4.7% on GSM8K). Importantly, DGA achieves SOTA with standard **Direct Sampling** at inference, avoiding costly test-time tree search and thus remaining practical under strict latency constraints.

## 6 Conclusion

We propose **Diffusion-Gibbs Alignment (DGA)**, which formulates RL for diffusion LLMs as variational distribution matching. A learned energy function approximates the optimal Gibbs policy, avoiding intractable marginals and enabling stable, gradient-free optimization. Across reasoning, math, and code, DGA achieves SOTA with better stability and efficiency than tree-search methods, offering a 3× speedup, **40%** less memory, and direct-sampling inference.

## 7 Limitations

First, while DGA eliminates the need for costly test-time tree search, training the auxiliary energy function introduces a modest computational overhead compared to simpler baselines like Diffu-GRPO. Second, our experimental evaluation is currently limited to the LLaDA-8B backbone on reasoning and coding tasks, leaving the scalability of DGA to larger parameter scales and open-ended generation domains for future investigation. Finally, although we employ weight clipping to ensure numerical stability, the reliance on self-normalized importance sampling may still face challenges with high variance if the proposal distribution diverges significantly from the optimal Gibbs target in highly complex search spaces.

## 8 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, and in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University under Grant 2025RGZN009.

## References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. 2021. Flow network based generative models for non-iterative diverse candidate generation. *Preprint*, arXiv:2106.04399.

Adam Block, Alexander Rakhlin, and Abhishek Shetty. 2024. On the performance of empirical risk minimization with smoothed data. *Preprint*, arXiv:2402.14987.

Mathieu Blondel, Michael E. Sander, Germain Vivier-Ardisson, Tianlin Liu, and Vincent Roulet. 2025. Autoregressive language models are secretly energy-based models: Insights into the lookahead capabilities of next-token prediction. *Preprint*, arXiv:2512.15605.

Christian Brownlees and Guđmundur Stefán Guđmundsson. 2023. Performance of empirical risk minimization for linear regression with dependent data. *Econometric Theory*, 41(2):391–420.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Deepseek. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Yilun Du, Jiayuan Mao, and Joshua B. Tenenbaum. 2024. Learning iterative reasoning through energy diffusion. *Preprint*, arXiv:2406.11179.

Yilun Du and Igor Mordatch. 2020. Implicit generation and generalization in energy-based models. *Preprint*, arXiv:1903.08689.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Susung Hong. 2024. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Preprint*, arXiv:2408.00760.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Vaibhav Jindal, Hejian Sang, Chun-Mao Lai, Yanning Chen, and Zhipeng Wang. 2025. Aligning diffusion language models via unpaired preference optimization. *Preprint*, arXiv:2510.23658.

Guillaume Lecué and Shahar Mendelson. 2016. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3).

Sergey Levine. 2018. Reinforcement learning and control as probabilistic inference: Tutorial and review. *Preprint*, arXiv:1805.00909.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025. A survey on diffusion language models. *Preprint*, arXiv:2508.10875.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. *Preprint*, arXiv:2210.02747.

llama team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *Preprint*, arXiv:2502.09992.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Zheyu Fu, Liancheng Fang, Minghua He, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. 2025. d-treerpo: Towards more reliable policy optimization for diffusion language models. *Preprint*, arXiv:2512.09675.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Yang Song and Diederik P. Kingma. 2021. How to train your energy-based models. *Preprint*, arXiv:2101.03288.

Xiaohang Tang, Rares Dolga, Sangwoong Yoon, and Ilija Bogunovic. 2025. wd1: Weighted policy optimization for reasoning in diffusion language models. *Preprint*, arXiv:2507.08838.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. llama2. *Preprint*, arXiv:2307.09288.

Shaoan Xie, Lingjing Kong, Xiangchen Song, Xinshuai Dong, Guangyi Chen, Eric P. Xing, and Kun Zhang. 2025. Step-aware policy optimization for reasoning in diffusion large language models. *Preprint*, arXiv:2510.01544.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. 2025. Energy-based diffusion language models for text generation. *Preprint*, arXiv:2410.21357.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025b. Mmada: Multimodal large diffusion language models. *Preprint*, arXiv:2505.15809.

Runpeng Yu, Qi Li, and Xinchao Wang. 2025. Discrete diffusion in large language and multimodal models: A survey. *Preprint*, arXiv:2506.13759.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. ReST-MCTS*: LLM self-training via process reward guided tree search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. 2022. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. 2025. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *Preprint*, arXiv:2504.12216.

Dingwei Zhu, Shihan Dou, Zhiheng Xi, Senjie Jin, Guoqiang Zhang, Jiazheng Zhang, Junjie Ye, Mingxu Chai, Enyu Zhou, Ming Zhang, Caishuang Huang, Yunke Zhang, Yuran Wang, and Tao Gui. 2025. Vrpo: Rethinking value modeling for robust rl training under noisy supervision. *Preprint*, arXiv:2508.03058.