DECODING LARGE LANGUAGE DIFFUSION MODELS WITH FORESEEING MOVEMENT

Yichuan Mo[1]* **Quan Chen**[2]* **Mingjie Li**[3] **Zeming Wei**[2] **Yisen Wang**[1,4]†

[1] State Key Lab of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[2] School of Mathematical Sciences, Peking University
[3] CISPA Helmholtz Center for Information Security, Germany
[4] Institute for Artificial Intelligence, Peking University

ABSTRACT

Large Language Diffusion Models (LLDMs) benefit from a flexible decoding mechanism that enables parallelized inference and controllable generations over autoregressive models. Yet such flexibility introduces a critical challenge: inference performance becomes highly sensitive to the decoding order of tokens. Existing heuristic methods, however, focus mainly on local effects while overlooking long-term impacts. To address this limitation, we propose the Foreseeing Decoding Method (FDM), a novel approach that integrates both local and global considerations to unlock the full potential, employing a search-based strategy to enable effective optimization in discrete spaces. Furthermore, by analyzing the consistency of chosen tokens in the full decoding process, we develop a variant, FDM with Acceleration (FDM-A), which restricts deep exploration to critical steps identified as the exploration and balance circumstances. Extensive experiments across diverse benchmarks and model architectures validate the scalability of FDM and demonstrate the superior efficiency-performance trade-off achieved by FDM-A. Our work might potentially provide a principled step toward more powerful decoding methods for LLDMs.

1 INTRODUCTION

In recent years, diffusion models (Ho et al., 2020; Rombach et al., 2022) have emerged as a promising competitor in natural language modeling (Nie et al., 2025; DeepMind, 2025), challenging the dominance of auto-regressive generation (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025). Instead of generating tokens sequentially from left to right and token by token, Large Language Diffusion Models (LLDMs) can generate multiple tokens in parallel, making it highly efficient at the inference stage (Li et al., 2025; Khanna et al., 2025). In addition, owing to pipelines of processing denoising and the bidirectional modeling, LLDMs can outperform the auto-regressive models in broad aspects such as reverse reasoning (Berglund et al., 2024), controllable generation (Xiong et al., 2025; Li et al., 2022), or multi-modal reasoning (Li et al., 2025).

Table 1: The performances of LLaDA with various decoding orders on the ARC (Clark et al., 2018) benchmark.

	Random	Margin	FDM-A
Accuracy (↑)	79.06	82.55	**86.30**
Tokens/Seconds (↑)	12.01	10.85	**38.20**

However, high flexibility is a double-edged sword, as it can lead to performance degradation if sampling paths are chosen inappropriately, attributing to the convergence issues (Li & Cai, 2025) and the non-convexity of the optimization goal (Kim et al., 2025; Peng et al., 2025; Ben-Hamu et al., 2025). For example, as shown in Table 1, compared to decoding with a random order, decoding answers with the order of the largest marginal probability yields an improvement on the ARC benchmark (Clark et al., 2018) from 79.06% to 82.55%, underscoring its critical importance. To address this issue, previous works have mainly developed strategies from a heuristic perspective. For

*Equal Contribution.
†Corresponding author: Yisen Wang (yisen.wang@pku.edu.cn).

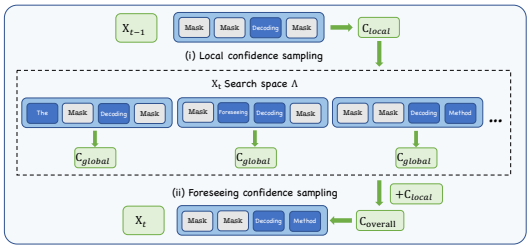


Figure 1: The pipeline of FDM. We first compress the search space into a small set Λ by filtering out candidates of lower local confidence *i.e.* C_{local}. In the final, we incorporate both local and global confidence to decide the ultimate choice at step t.

example, pioneer works (Nie et al., 2025; Zheng et al., 2024) propose to decode by enlarging the predicted probability. Follow-up works argue that probability margins (Kim et al., 2025) and the entropy of the predicted distribution (Ben-Hamu et al., 2025) are better substitutes, considering that LLDMs may be confused when the probabilities of the topmost tokens at the given position are near. However, though effective, previous heuristic approaches decode tokens with local information at each step, overlooking the long-term impacts underlying the full decoding paths without leveraging the full potentiality of LLDMs at the inference stage.

Therefore, in this paper, with further analyzing the decoding formulation of LLDMs, we find that it can be divided into two components. The first one is termed as **local confidence** in Figure 1 (i), reflecting the uncertainty of models in predicting a given token and widely adopted by the heuristic methods. The other is the **global confidence**, capturing the future impact in decoding a specific token. It is not directly accessible and typically ignored by prior methods. However, we demonstrate that the training goal of LLDMs provides a good approximation. By further combining both of them into considerations, we propose **F**oreseeing **D**ecoding **M**ethod (FDM), illustrated in Figure 1. To ensure the computation is affordable in reality, the decoding process of FDM is divided into two stages: We firstly rank the candidate tokens with their local confidence. In the second competition, the overall criterion of both local and global confidence is adopted to decide the final tokens chosen for decoding. We not only theoretically prove FDM will achieve lower errors compared to heuristic methods but also perform experiments across multiple benchmarks to illustrate that the benefit of it can further improve by enlarging the search space. It demonstrates FDM will serve as a test-time scaling (Snell et al., 2024; Bi et al., 2024) method specifically designed for LLDMs.

Furthermore, by calculating the consistency ratio of FDM and the simple confidence-based decoding, we find that a global view is not always necessary at every step, especially when the context is sufficient for reliable decisions. Motivated by this observation, we propose an accelerated version of FDM (FDM-A). It will perform an adaptive exploration from the probability feedback, switching to FDM only when all candidates have low confidence or borderline scenarios. As shown in Table 1, FDM-A can not only accelerate the decoding process with over $3\times$ speed-up but also obtain notable performances. In summary, our contributions are listed as follows:

- We propose the **F**oreseeing **D**ecoding **M**ethod (FDM) to schedule the decoding orders of LLDMs. Unlike heuristic approaches, it takes both the local and global confidence as the criterion for decision, guaranteeing a lower divergence with the natural distribution.
- Motivated by the agreement ratio in decoding, we further accelerate FDM with an adaptive strategy (FDM-A). It adaptively performs exploration when the context is rare and parallel decodes tokens if models have enough context for complete generation.
- The experiments across multiple benchmarks and variants of LLDMs are performed, manifesting the scalability of FDM and the outstanding performance of FDM-A in balancing efficiency and performance.

2 Related Work

Large Language Diffusion Models (LLDMs). As one of the most successful types of generative models, the popularity of diffusion models has been well-recognized in the vision domain (Ho et al., 2020; Peebles & Xie, 2023; Rombach et al., 2022). At the training stage, the model predicts

the noises added to the natural inputs. Then, at the test time, by utilizing the model to denoise step by step, diffusion models recover the original data through iterative refinement. Inspired by their great success, early investigations such as D3PMs (Austin et al., 2021), DiffuSeq (Gong et al., 2023) have demonstrated the potential of the diffusion pipeline in language tasks. However, limited to the small scales, their performances are considerably inferior to those of their auto-regressive competitors. Recently, LLaDA (Nie et al., 2025; Zhu et al., 2025a) scales the model parameters to billions, considered as one of the most representative works in Large Language Diffusion Models (LLDMs). To save the computational costs, other architectures like Dream (Ye et al., 2025), DiffuGPT (Gong et al., 2024), and Dream-Coder (Xie et al., 2025) perform continual training on the pretrained weights of the auto-regressive counterparts. Built upon the outstanding performances in the language capacity, large language diffusion models have also been applied to multi-modal applications such as biomedical understanding (Dong et al., 2025), chart understanding (You et al., 2025), and mathematical reasoning (Yang et al., 2025). But in this paper, we propose FDM to improve the performance of LLDMs at the inference stage.

Decoding Order of LLDMs. With the rapid development of LLDMs, their shortcomings are also disclosed. One of the most prominent ones is the vital significance of the decoding orders (Kim et al., 2025). This is because when the context is limited, models will fail to accurately predict all tokens. In other words, the uncertainty of predictions can be measured with the output probability. This motivates the heuristic-based approaches, including decoding with the largest probability value (Nie et al., 2025; Zheng et al., 2024), with the largest margin probability (Kim et al., 2025), and with the least entropy of distribution (Ben-Hamu et al., 2025). Although heuristic methods improve the performance of LLDMs a lot compared to decoding with the random order, they make decisions based on local confidence, ignoring the sequential consequence to other tokens. This limitation may bring errors to the generated sequences. Thus, in this paper, we propose FDM and FDM-A, which incorporate the global confidence to depict the future cascading effect in the decoding process. .

Acceleration of LLDMs. Although LLDMs map to the complete answer at each step, they will suffer a quality-efficiency trade-off with a varying decoding step (Feng et al., 2025; Nie et al., 2025). In addition, to achieve the bidirectional attentions, the attention maps of LLDMs are not causal, requiring model-agnostic methods like KV-Cache for adaptations. Therefore, in a series of works (Wu et al., 2025; Hu et al., 2025; Ma et al., 2025), they propose their own approximation acceleration for reusing the Key-Value, largely accelerating the inference speed. More related to our work, Hong et al. (2025) and Ben-Hamu et al. (2025) both propose samplers for dynamic decoding in LLDMs. But Hong et al. (2025) purely focus on the parallel decoding for acceleration and Ben-Hamu et al. (2025) propose WINO for revoking suspicious error tokens. While in our paper, we propose FDM-A to balance the explorations and accelerations in decoding, achieving a better trade-off.

3 PRELIMINARIES

Unlike auto-regressive models that generate answers in a sequential way, LLDMs define a Markov chain that revises the answer step-by-step with a denoised process. Given the user query $\mathbf{q}$ and the vocabulary space $M = \{1, 2, ...m\}$, the generation of final answer $\mathbf{x}_T$ with length L under the data distribution p_{data} can be formulated as:

$$\mathbf{x}_T = \arg\max_{\mathbf{x}_T} p_{data}(\mathbf{x}_{0:T}|\mathbf{q}) = \arg\max_{\mathbf{x}_T} \{p(\mathbf{x}_0) \prod_{\alpha=1}^{T-1} p_{data}(\mathbf{x}_\alpha|\mathbf{q}, \mathbf{x}_{0:\alpha-1})\}, \tag{1}$$

where $p(\mathbf{x}_0)$ is sampled from an initial noise distribution which is independent to $\mathbf{q}$. $\mathbf{x}_\alpha$ is a partially masked sequence at the α step. To analyze the influence of the intermediate variable $\mathbf{x}_t$, we can regroup the terms associated with step t as:

$$\mathbf{x}_T = \arg\max_{\mathbf{x}_T} \{p_{data}(\mathbf{x}_{t+1:T}|\mathbf{q}, \mathbf{x}_{0:t}) * p_{data}(\mathbf{x}_t|\mathbf{q}, \mathbf{x}_{t-1}) \prod_{\alpha=1}^{t-1} p_{data}(\mathbf{x}_\alpha|\mathbf{q}, \mathbf{x}_{0:\alpha-1})\}. \tag{2}$$

Although the above equation shows that $\mathbf{x}_t$ can impact future variables from $\mathbf{x}_{t+1}$ to $\mathbf{x}_T$, heuristic decoding methods commonly simplify the procedure by focusing solely with the local partition:

$$\pi_H(\mathbf{x}_t|\mathbf{x}_{t-1}) : \mathbf{x}_t = \arg\max_{\mathbf{x}_t} p_\theta(\mathbf{x}_t|\mathbf{q}, \mathbf{x}_{t-1}), \tag{3}$$

where θ is a parameterized neural network introduced by LLDMs to model the natural distribution. According to (Nie et al., 2025), the optimization target of LLDMs can be given as:

$$\mathbb{E}_{\substack{\mathbf{x}_T\sim p_{data}\\ t\in[0,T]}}\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[\mathbf{x}_t^{(j)}=\texttt{Mask}]\odot\log p_\theta(\mathbf{q},\mathbf{x}_t^{(j)})[\mathbf{x}_T], \tag{4}$$

where $\odot$ is the Hadamard product and $[\mathbf{x}_T]$ represents the probability value of the tokens in $\mathbf{x}_T$. With a well-trained model, LLDMs learn to approximate the conditional log-probability of future tokens in $\mathbf{x}_m$ given the current masked state, $\mathbf{x}_t$ ($m > t$) and the user query $\mathbf{q}$. Specifically,

$$\log p_\theta(\mathbf{x}_m|\mathbf{x}_t,\mathbf{q})=\mathbf{1}[\mathbf{x}_m\neq\texttt{Mask}\ \&\&\ \mathbf{x}_t=\texttt{Mask}]\odot\log p_\theta(\mathbf{q},\mathbf{x}_t)[\mathbf{x}_m], \tag{5}$$

which follows from the fact that LLDMs predict the tokens for any masked position independently.

4 METHODOLOGY

4.1 THE FORESEEING DECODING METHOD

Although Eq. 3 establishes a simple formulation in decoding $\mathbf{x}_t$, ignoring the contribution of $\mathbf{x}_t$ in $p_{data}(\mathbf{x}_{t+1:T}|\mathbf{q},\mathbf{x}_{0:t})$ might mislead the decoding process, leading the unexpected errors in the generative response. Therefore, differing from previous works, our proposed Foreseeing Decoding Method (FDM) aims to decode tokens that achieve the largest value across the overall formulation and its decoding strategy can be given as:

$$\pi_F(\mathbf{x}_t|\mathbf{q},\mathbf{x}_{t-1}):\ \mathbf{x}_t=\arg\max_{\mathbf{x}_t}p_\theta(\mathbf{x}_T|\mathbf{q},\mathbf{x}_t)\,p_\theta(\mathbf{x}_t|\mathbf{q},\mathbf{x}_{t-1}). \tag{6}$$

Here we replace $p_{data}(\mathbf{x}_{t+1:T}|\mathbf{q},\mathbf{x}_{0:t})$ with $p_\theta(\mathbf{x}_T|\mathbf{q},\mathbf{x}_t)$ owing to the property of the Markov Chain and the modeling of LLDMs. From the theoretical aspect, we also prove that under π_F, the generative distribution achieves lower KL divergence with the natural distribution p_{data} than that of π_H.

Theorem 1. *Let* $\Delta_{total}\triangleq\sum_{t=1}^{T}\mathbb{E}_{p_{data}(\mathbf{x}_{t-1})}\big[\mathcal{I}_{p_{data}}(\mathbf{x}_t;\mathbf{x}_T|\mathbf{x}_{t-1})\big]$, *where* $\mathcal{I}_{p_{data}}(\mathbf{x}_t;\mathbf{x}_T|\mathbf{x}_{t-1})$ *is the conditional mutual information under* q. *Then*

$$D_{KL}(p_{data}(\mathbf{x}),p_{\pi_F})=D_{KL}(p_{data}(\mathbf{x}),p_{\pi_H})-\Delta_{total}. \tag{7}$$

For the complete proof, please refer to Appendix B for more details. Due to the fact that mutual information is non-negative, we derive that the generative distribution under π_F has lower KL divergence with the natural distribution p_{data} than π_H. By further applying the log transformation to Eq. 6, we can obtain,

$$\mathbf{x}_t=\arg\max_{\mathbf{x}_t}\left\{\log p_\theta(\mathbf{x}_T|\mathbf{q},\mathbf{x}_t)+\log p_\theta(\mathbf{x}_t|\mathbf{q},\mathbf{x}_{t-1})\right\}. \tag{8}$$

With deeper analysis, we find the first term captures how decoding a specific token influences the future (the **global** confidence, C_{global}), while the second term reflects the model's confidence in decoding it at step t (the **local** confidence, C_{local}). According to Eq.5, C_{global} can be estimated by:

$$C_{global}=\log p_\theta(\mathbf{x}_T|\mathbf{x}_t,\mathbf{q})=\mathbf{1}[\mathbf{x}_T\neq\texttt{Mask}\ \&\&\ \mathbf{x}_t=\texttt{Mask}]\odot\log p_\theta(\mathbf{q},\mathbf{x}_t)[\mathbf{x}_T]. \tag{9}$$

Regarding $\mathbf{x}_T$ is a full response without any mask, $\mathbf{x}_T\neq\texttt{Mask}$ can be omitted since it holds for every position. In addition, instead of greedily decoding $\mathbf{x}_T$ from the model output to calculate C_{global}. We compute the expectation over the entire model output distribution:

$$C_{global}=\mathbf{1}[\mathbf{x}_t=\texttt{Mask}]\odot\mathbb{E}_{p_\theta}\log p_\theta(\mathbf{q},\mathbf{x}_t). \tag{10}$$

In addition, with Eq. 5, we can formulate C_{local} with:

$$C_{local}=\mathbf{1}[\mathbf{x}_t\neq\texttt{Mask}\ \&\&\ \mathbf{x}_{t-1}=\texttt{Mask}]\odot\log p_\theta(\mathbf{q},\mathbf{x}_{t-1})[\mathbf{x}_t]. \tag{11}$$

Similar to the denotation in Eq. 5, here $[\mathbf{x}_t]$ means the predicted probability of each token in $\mathbf{x}_t$. After combining both equations, we have:

$$\begin{aligned}\mathbf{x}_t=\arg\max_{\mathbf{x}_t}\{&\mathbf{1}[\mathbf{x}_t=\texttt{Mask}]\odot\mathbb{E}_{p_\theta}\log p_\theta(\mathbf{q},\mathbf{x}_t)\\&+\mathbf{1}[\mathbf{x}_t\neq\texttt{Mask}\ \&\&\ \mathbf{x}_{t-1}=\texttt{Mask}]\odot\log p_\theta(\mathbf{q},\mathbf{x}_{t-1})[\mathbf{x}_t]\}.\end{aligned} \tag{12}$$

Note that $p_\theta(\mathbf{q},\mathbf{x}_{t-1})$ is independent of $\mathbf{x}_t$, which can be accurately calculated by querying the model θ with the acquired sequence $\mathbf{x}_{t-1}$ and user prompt $\mathbf{q}$. In contrast, the $p_\theta(\mathbf{q},\mathbf{x}_t)$ takes the discrete

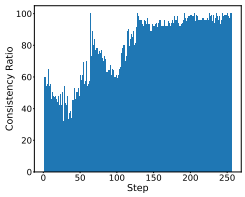


Figure 2: The consistency ratio of selecting the next decoding token using C_{local} versus both C_{local} and C_{global}.

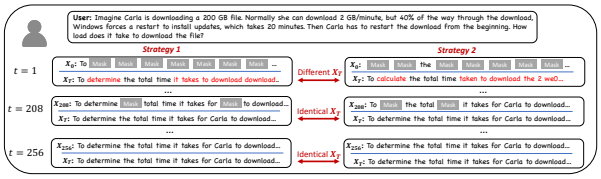


Figure 3: The effect of different decoding strategy to $\mathbf{x}_T$ at the step t given the identical $\mathbf{x}_{t-1}$. The influence gradually decreases as t increases from 0 to T.

variable $\mathbf{x}_t$ as a part of the input. It means every evaluation involves a single forward pass. To resolve this problem, we introduce the hyperparameter K to compress the search space for efficiency. In detail, we first obtain the set of candidate tokens with the prediction of θ on the masked position:

$$\texttt{Candidate} = \left\{\mathbf{1}[\mathbf{x}_{t-1}{=}\texttt{Mask}] \odot \arg\max_{\mathbf{x}} p_\theta(\mathbf{q}, \mathbf{x}_{t-1})\right\}. \tag{13}$$

Furthermore, to avoid being trapped in the local optima, we also incorporate a dynamic pruning strategy that retains only candidate tokens whose confidence exceeds the predefined threshold γ. Then, C_{local} is introduced as the metric to rank each possibility of $\mathbf{x}_t$ and we only keep the Top-K candidates for further decision. Thus, the search space Λ can be defined as:

$$\Lambda = \{\mathbf{x}_t | \mathbf{x}_t \in \text{Top-K}(\{\mathbf{x}_t\}),\ \mathbf{1}[\mathbf{x}_t \neq \texttt{Mask}\ \&\&\ \mathbf{x}_{t-1} = \texttt{Mask}] \odot p_\theta(\mathbf{q}, \mathbf{x}_{t-1})[\mathbf{x}_t] > \gamma\}. \tag{14}$$

Defined on Λ, our proposed **F**oreseeing **D**ecoding **M**ethod (FDM) is formulated as follows,

$$\mathbf{x}_t = \begin{cases} \arg\max_{\mathbf{x}_t \in \{\mathbf{x}_t\}} C_{local}(\mathbf{x}_t) & \text{if } \Lambda = \varnothing \\ \arg\max_{\mathbf{x}_t \in \Lambda} \{C_{local}(\mathbf{x}_t) + C_{global}(\mathbf{x}_t)\} & \text{if } \Lambda \neq \varnothing \end{cases} \tag{15}$$

We also summarize the whole decoding process of FDM at the t step in Algorithm 1 in Appendix C.

4.2 Acceleration with the Foreseeing Decoding Method

Based on the analysis in Section 4.1, in this section, we further investigate a variant, *i.e.*, FDM-A to better balance the decoding speed and performance. In Fig. 2, we calculate the consistency ratio in token selection with local confidence only (heuristic decoding) versus both local and global confidence. The decisions of both strategies are made based on the same $\mathbf{x}_{t-1}$ in each step. We average the results over the first 100 examples in the GSM8K (Cobbe et al., 2021) test set. The experiments are performed on LLaDA (Nie et al., 2025), with $\gamma = 0.6$ and $K = 2$. Peak points are observed on the steps of 64 and 128 because we follow the proposed semi-autoregressive pipeline in (Nie et al., 2025) with the block size 64.

In the early stages of decoding, due to limited context, the overlap between the two decoding strategies is relatively low, at approximately 50%. However, as decoding progresses, the degree of overlap gradually increases and finally exceeds 90%. It is consistent with the empirical observation in Fig. 3, the final decoding sequence ($\mathbf{X}_T$) is largely shaped when t is small. But when t is approaching T, most unmasked tokens have been determined by the adequate contexts, indicating the marginal impact of decoding orders. This evidence demonstrates that more exploration from the global confidence is needed at the beginning, while at the final stage, we only need to adopt the local confidence to accelerate the decoding.

Motivated by this observation, we divide the entire process into three stages: **the exploration, balance and acceleration phases** by incorporating two hyperparameters, *i.e.* η_1 and η_2. Firstly, in the exploration phase, if there are no tokens whose prediction probability exceeds η_1, we will perform exploration with FDM. We denote it with $\texttt{FDM}_1$ $(\mathbf{x}_{t-1}, \mathbf{q}, n)$, initialized with the K_1. n is the number of tokens decoded in this step and we set it to 1 for reliability. Thus $\mathbf{x}_t$ can be given as:

$$\texttt{FDM}_1(\mathbf{x}_{t-1}, \mathbf{q}, n = 1, \gamma = \gamma_1). \tag{16}$$

In the balance phase, we focus on the intermediate states of the decoding process. We combine the exploration and acceleration to achieve a good trade-off. Specifically, we refer to the tokens whose probability lies between η_1 and η_2 as the "Borderline Tokens" and those that have the

Table 2: Comparison (%) of FDM with the heuristic methods. With the scale up of K, accuracy increases with the decrease of TPS, illustrating that FDM serves as an inference-time scaling method.

Benchmark	Method	LLaDA		LLaDA-1.5		MMaDA-MixCoT		LLaDA-MoE	
		Accuracy	TPS	Accuracy	TPS	Accuracy	TPS	Accuracy	TPS
GSM8K	Probability (T=256)	81.20	11.51	82.10	11.89	56.18	11.26	76.50	4.13
	Margin (T=256)	80.14	11.23	82.18	11.39	56.18	10.97	76.80	4.04
	Entropy (T=256)	80.12	10.79	81.93	10.94	54.91	10.51	76.80	3.84
	FDM (K=2)	82.03	8.28	82.49	8.29	57.54	8.18	77.48	3.80
	FDM (K=3)	82.18	5.86	82.64	5.84	57.68	5.71	77.63	3.78
	FDM (K=4)	82.34	4.61	83.02	4.69	57.92	4.61	78.32	3.60
HumanEval	Probability (T=256)	42.68	9.24	42.68	9.20	12.80	8.91	56.71	3.81
	Margin (T=256)	43.29	8.76	42.68	8.83	12.80	8.65	58.54	3.81
	Entropy (T=256)	40.24	8.67	42.68	8.67	11.59	8.28	59.15	3.56
	FDM (K=2)	43.29	6.55	43.29	6.66	12.80	6.32	59.76	3.72
	FDM (K=3)	44.51	4.63	43.29	4.51	12.80	4.60	59.76	3.67
	FDM (K=4)	45.73	3.66	44.51	3.71	13.40	3.64	60.37	3.47
Countdown	Probability (T=256)	18.75	11.19	16.02	11.20	4.69	11.15	40.62	4.15
	Margin (T=256)	19.14	10.89	20.31	10.92	14.61	10.84	40.23	4.03
	Entropy (T=256)	18.44	10.39	20.31	10.39	7.03	10.31	38.67	3.95
	FDM (K=2)	19.14	8.21	20.70	8.62	14.45	8.13	40.62	3.73
	FDM (K=3)	21.09	6.11	21.88	6.48	16.40	6.10	46.10	3.71
	FDM (K=4)	25.00	5.28	23.43	5.25	17.58	4.97	46.88	3.70
ARC	Probability (T=256)	80.96	10.96	87.06	10.98	58.26	10.94	76.19	4.01
	Margin (T=256)	82.55	10.85	87.44	10.82	56.37	10.62	75.85	4.01
	Entropy (T=256)	78.13	10.43	87.38	10.38	58.33	10.18	71.84	3.98
	FDM (K=2)	86.00	7.72	88.18	7.71	59.43	7.68	82.89	3.85
	FDM (K=3)	86.46	5.58	88.45	5.59	59.61	5.58	83.45	3.73
	FDM (K=4)	86.68	4.58	88.59	4.56	59.76	4.37	83.51	3.64

prediction probability greater than η_1 as the "Qualified Tokens". By applying FDM, we weigh options within a collection comprising both sets. The number of decoding tokens, n is set as $\texttt{NUM}\big([\mathbf{x}_{t-1} = \texttt{Mask}]p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1\big)$ for acceleration and $\texttt{NUM}(\cdot)$ is a function that counts the number of tokens that meet the given condition. The formulation of decoding $\mathbf{x}_t$ is:

$$\texttt{FDM}_1\Big(\mathbf{x}_{t-1}, \mathbf{q},\ n = \texttt{NUM}\big([\mathbf{x}_{t-1} = \texttt{Mask}]\, p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1\big),\ \gamma = \eta_2\Big). \tag{17}$$

As shown in Fig. 3, different decoding methods perform almost the same at the tail, indicating the convergence in decoding. At this time, we shift the mode to the acceleration phase. The tokens are decoded with the local confidence only to achieve the fastest speed. The decoding function can be regarded as a special case of FDM. By initializing it with $K = 1$, we represent it by $\texttt{FDM}_2(\mathbf{x}_{t-1}, \mathbf{q}, n)$ and $\mathbf{x}_t$ can be decoded with:

$$\texttt{FDM}_2\Big(\mathbf{x}_{t-1}, \mathbf{q},\ n = \min\big(\texttt{NUM}\big([\mathbf{x}_{t-1} = \texttt{Mask}]\, p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1\big), N\big),\ \gamma = 1.0\Big). \tag{18}$$

N is the upper bound for clipping the decoding number of tokens. We term this accelerated decoding pipeline as FDM-A. The algorithm of it is summarized in Algorithm 2 in Appendix D.

5 Experiment

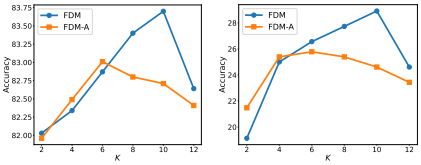
5.1 Main Results

Benchmarks and Metrics: To demonstrate the effectiveness of FDM and FDM-A, we conduct experiments on four prevailing benchmark datasets: GSM8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021), Countdown (Zhao et al., 2025) and ARC (Clark et al., 2018). They reflect the capabilities of LLDMs in four aspects, including math issue solutions, code generation, logic reasoning, and common sense knowledge. Meanwhile, for the convenience of evaluation, following (Hong et al., 2025), we add a system prompt before each input item. The details of them in each dataset are summarized in Appendix E. All evaluations are performed under the zero-shot condition to ensure the fairness. For performances, we take accuracy as the metric, manifesting the percentage of queries, that could be properly answered by LLDMs. For efficiency, we choose Tokens Per Second (TPS) as the metric. It is defined as the number of tokens generated in one second.

Baselines and models: We compare FDM with three heuristic-based decoding methods: decoding with the highest probability (Probability) (Nie et al., 2025; Zheng et al., 2024), with the highest marginal probability (Margin) (Kim et al., 2025) and with the lowest entropy (Entropy) (Ben-Hamu et al., 2025). For FDM-A, since it tries to balance between efficiency and performance, we also compare it with the dynamic decoding methods for LLDMs: EB (Entropy Bounded Sampler) (Ben-Hamu et al., 2025) and WINO (Hong et al., 2025). To ensure the generalability of the observations, we cover four variants of LLDMs: LLaDA-8B-Instruct (Nie et al., 2025), LLaDA-1.5 (Zhu et al., 2025a), LLaDA-MoE-7B-Instruct (Zhu et al., 2025b) and MMaDA-8B-MixCoT (Yang et al., 2025).

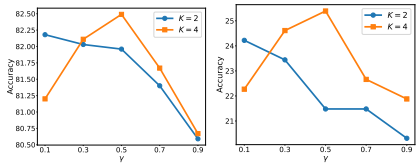
Table 3: Comparison (%) of FDM-A with methods that accelerate the decoding of LLDMs. The best results are in **bold**. FDM-A achieves not only the highest accuracy, but also the fastest speed.

Benchmark	Method	LLaDA		LLaDA-1.5		MMaDA-MixCoT		LLaDA-MoE	
		Accuracy	TPS	Accuracy	TPS	Accuracy	TPS	Accuracy	TPS
GSM8K	Probability (T=128)	78.17	20.86	80.06	22.04	52.08	21.65	71.49	8.81
	Margin (T=128)	79.23	21.36	80.39	21.61	50.72	21.37	72.93	8.47
	Entropy (T=128)	77.94	20.69	80.22	20.71	49.13	20.51	70.66	8.39
	EB	79.61	36.46	81.35	37.74	53.83	36.39	74.53	13.98
	WINO	79.45	40.02	80.89	40.96	51.33	36.92	72.78	13.02
	FDM-A (Ours)	**81.96**	**42.65**	**82.87**	**43.98**	**55.12**	**40.96**	**76.72**	**14.89**
HumanEval	Probability (T=128)	35.37	16.75	39.63	17.04	7.93	16.59	51.22	7.99
	Margin (T=128)	37.20	16.25	37.20	16.58	12.80	16.15	49.39	7.74
	Entropy (T=128)	31.20	15.93	35.37	15.77	4.27	15.46	48.78	7.46
	EB	37.20	17.34	37.80	18.99	12.80	24.95	59.15	10.14
	WINO	39.02	18.33	40.24	20.59	12.20	29.90	56.71	9.31
	FDM-A (Ours)	**44.51**	**21.56**	**42.07**	**23.83**	**13.41**	**32.32**	**60.98**	**11.25**
Countdown	Probability (T=128)	19.53	21.19	**21.09**	22.01	12.01	21.97	44.53	8.80
	Margin (T=128)	20.31	20.65	19.92	21.76	8.59	21.57	42.58	8.33
	Entropy (T=128)	19.53	20.62	18.75	20.93	8.59	20.52	35.94	8.15
	EB	19.20	18.68	19.92	18.79	18.75	48.51	42.19	9.48
	WINO	19.14	20.52	19.14	20.38	20.70	52.39	25.00	8.23
	FDM-A (Ours)	**21.48**	**21.98**	**21.09**	**22.29**	**26.95**	**62.30**	**49.61**	**10.14**
ARC	Probability (T=128)	82.16	21.24	86.38	21.71	55.04	22.26	72.74	8.68
	Margin (T=128)	84.34	20.65	85.91	21.34	56.30	21.47	75.48	8.56
	Entropy (T=128)	77.20	19.91	84.58	20.48	55.61	20.58	69.08	8.12
	EB	73.55	32.01	85.68	34.89	56.86	32.89	71.50	8.98
	WINO	79.44	34.17	85.31	35.27	56.79	34.27	70.86	7.92
	FDM-A (Ours)	**86.30**	**38.20**	**87.66**	**38.43**	**59.50**	**37.07**	**83.33**	**12.62**



(a) GSM8K (b) Countdown

Figure 4: The influence of K on model performance on GSM8K and Countdown benchmarks.



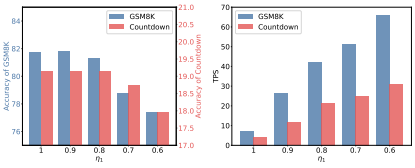
(a) GSM8K (b) Countdown

Figure 5: The influence of γ on model performance on GSM8K and Countdown benchmarks.

Hyperparameters: Consistent with the observation in (Nie et al., 2025), we observe that the semi-autoregressive pipeline is vital to maintain a satisfying performance for the instruct models of LLDMs. Thus we applying it for decoding with the generation length 256, block size 64. The step number T is fixed to 256 and 128, respectively for the heuristic decoding methods when comparing them with FDM or FDM-A. The threshold in EB is set as 0.5. τ_1 and τ_2 in WINO are configured as 0.7 and 0.9. For our method, γ for the dynamic pruning is 0.6 for FDM and 0.5 for FDM-A considering explorations are less performed in FDM-A. For the threshold for stage division, we set η_1 to 0.8 and η_2 to 0.7. We gradually increase the width of FDM from 2 to 4 and set the default width of FDM-A as 2. All experiments are performed on a single NVIDIA A100 80G GPU.

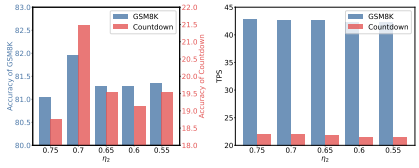
Results: Firstly, in Table 2, we observe that FDM surpasses baseline methods across all models and benchmarks, demonstrating the significance of exploration with the global confidence. For example, on the LLaDA model and ARC benchmarks, the highest score of the heuristic methods is 82.55%, and FDM with the search width 2 is 86.00%. In addition, scaling up the width K can further enhance its performance, demonstrating its role in serving as an inference-time scaling technique. An example is that on the LLaDA-MoE and GSM8K benchmark, the accuracy improves from 77.48% to 78.32% when the width rises from 2 to 4. It is worth noting that this improvement is notable since we do not need a verifier or optimize the parameters. It demonstrates that FDM well fits the scenarios that have sufficient computations but need high accuracy. In Appendix F, we show an example case that could be properly decoded with FDM but incorrectly decoded for all heuristic methods.

Secondly, in Table 3, the results reveal that FDM-A achieves an outstanding trade-off between accuracy and speed: When we compare FDM-A's performance with FDM, the negative influence is marginal. The accuracy of FDM ($K = 2$) on the GSM8K benchmark of the LLaDA model is 82.03%, and for FDM-A, it achieves 81.96% (-0.07%). Meanwhile, FDM-A achieves 5.15× speedup in TPS, significantly improving TPS from 8.28 to 42.65. In contrast, the performance of heuristic methods will largely degrade when their decoding step is reduced. For example, when we compare the results in Table 2 and 3, the accuracy with the highest probability decoding declines from 81.20% to 78.17% when we halve the step number. Owing to its full exploitation with the global confidence in the exploration and balance stage, FDM-A also outperforms dynamic decoding methods like WINO.



(a) Accuracy (b) TPS

Figure 6: The influence of η_1 on Accuracy and TPS on GSM8K and Countdown benchmarks.



(a) Accuracy (b) TPS

Figure 7: The influence of η_2 on Accuracy and TPS on GSM8K and Countdown benchmarks.

5.2 Ablation Studies

Going Wider with K. In Section 5.1, we show that larger K will benefit the performance of FDM and a natural question is whether it will bring consistent improvement with the increase of K. Thus, here we set K with the value from $[2, 4, 6, 8, 10, 12]$ and perform experiments on the LLaDA model. As summarized in both Fig. 4 and Fig. 8 in Appendix G, the accuracy will reach the peak for both FDM-A and FDM. This is because although the training goal of the network is to fit the data distribution p_{data}, in realistic, the divergence between two distributions is not zero, leading to accumulating bias when we perform excessive searching operations, supported by the theoretical analysis in Appendix H. Intriguingly, when we compare the curves of FDM and FDM-A, we find that FDM outperforms FDM-A when K is larger. Conversely, FDM-A holds an advantage when K is small. This highlights the complementary role of both methods. FDM-A is suited for use when computational resources are limited. FDM, on the other hand, provides robust capabilities under conditions of abundant computational supply.

The Setting of γ. When we foresee the future impact with the global confidence, we introduce a threshold γ to dynamically prune the candidate tokens of the low local confidence value. We further investigate its effect by tuning it between 0.1 and 0.9 with $K = 2$ and 4. Taking FDM-A as an example, the results on LLaDA of the GSM8K and Countdown datasets are shown in Fig. 5. For more results, please refer to Fig. 9 in Appendix G. We also observe a trade-off in its configuration: the results illustrate that a smaller γ, *e.g.* 0.1, achieves better accuracy in $K = 2$. This is because it provides a wider choice for selection, mitigating the issue of inadequate search. However, we find that it can not maintain its advantage when K increases to 4, since more tokens of lower C_{local} are chosen, reflecting the uncertainty of models in its prediction. In contrast, if we set γ too large, it will suppress the exploration in the foreseeing decoding. In total, γ near 0.5 can balance both in its application, which is chosen as the configuration for the experiments in Section 5.1.

The Choices of η_1 and η_2. In addition to K and γ, FDM-A also introduces η_1 and η_2 as hyperparameters for stage divisions. In the main text, we perform experiments on the GSM8K and Countdown benchmarks in Fig. 6 and 7. For more results on HumanEval and ARC benchmarks, please refer to Fig. 10 and 11 in Appendix G. We firstly fix η_2 at 0.6 and linearly decrease η_1 from 1.0 to 0.6. The results in Fig. 6 (a) demonstrate that the accuracy remains stable at first, and then drops sharply in small η_1. In contrast, the TPS monotonically improves with the decrease of η_1 because more tokens are parallely decoded. By default, we set η_1 as 0.8 because it can not only maintain the model utility but also achieve high decoding speed. Secondly, we fix η_1 with 0.8 and set η_2 from [0.75, 0.7, 0.65, 0.6, 0.55]. Among all configurations, $\eta_2 = 0.7$ consistently achieves outstanding performances across all benchmarks. Because when η_2 is large, the exploration at the balance stage is insufficient. But if η_2 is set too small, uncertain tokens will interfere the correctness of decoding.

6 Conclusion

In this paper, we tackle the challenge of token decoding orders in Large Language Diffusion Models (LLDMs). We highlight the shortcomings of existing methods based on local confidence and propose a new approach by integrating global confidence. We prove that our method achieves lower divergence from the natural distribution. Our Foreseeing Decoding Method (FDM) optimizes decoding order via a heuristic beam search, balancing local and global factors. The accelerated version, FDM-A, boosts efficiency by applying deep exploration at critical steps. Experiments show that FDM outperforms current baselines, while FDM-A offers a strong performance-speed trade-off. Our work lays the foundation for more efficient LLDM decoding strategies.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *arXiv preprint arXiv:2505.24857*, 2025.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *ICLR*, 2024.

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

Google DeepMind. Gemini diffusion, 2025. URL https://deepmind.google/models/gemini-diffusion/.

Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. Llada-medv: Exploring large language diffusion models for biomedical image understanding. *arXiv preprint arXiv:2508.01617*, 2025.

Guhao Feng, Yihan Geng, Jian Guan, Wei Wu, Liwei Wang, and Di He. Theoretical benefit and limitation of diffusion language model. *arXiv preprint arXiv:2502.09622*, 2025.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In *ICLR*, 2023.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Feng Hong, Geng Yu, Yushi Ye, Haicheng Huang, Huangjie Zheng, Ya Zhang, Yanfeng Wang, and Jiangchao Yao. Wide-in, narrow-out: Revokable decoding for efficient and effective dllms. *arXiv preprint arXiv:2507.18578*, 2025.

Zhanqiu Hu, Jian Meng, Yash Akhauri, Mohamed S Abdelfattah, Jae-sun Seo, Zhiru Zhang, and Udit Gupta. Accelerating diffusion language model inference via efficient kv caching and guided diffusion. *arXiv preprint arXiv:2505.21467*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 2025.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *ICML*, 2025.

Gen Li and Changxiao Cai. A convergence theory for diffusion language models: An information-theoretic perspective. *arXiv preprint arXiv:2505.21400*, 2025.

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

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. In *NeurIPS*, 2022.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*, 2025.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling. *arXiv preprint arXiv:2502.03540*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025.

Zhihui Xie, Jiacheng Ye, Lin Zheng, Jiahui Gao, Jingwei Dong, Zirui Wu, Xueliang Zhao, Shansan Gong, Xin Jiang, Zhenguo Li, et al. Dream-coder 7b: An open diffusion language model for code. *arXiv preprint arXiv:2509.01142*, 2025.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Unveiling the potential of diffusion large language model in controllable generation. *arXiv preprint arXiv:2507.04504*, 2025.

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

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

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

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *COLM*, 2024.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, et al. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*, 2025a.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, et al. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*, 2025b.

A USAGE OF LLM

In this paper, we employ Large Language Models (LLMs) to polish writing and assist in debugging. All LLM-polished text is rechecked by the authors to ensure accuracy and prevent over-claims or confusion.

B PROOF OF THEOREM 1

Define the single-step KL errors

$$\varepsilon_t^H \triangleq D_{KL}(p_{data}(\mathbf{x}_t|\mathbf{x}_{t-1}), \pi_H(\mathbf{x}_t)), \qquad \varepsilon_t^F \triangleq D_{KL}(p_{data}(\mathbf{x}_t|\mathbf{x}_{t-1}), \pi_F(\mathbf{x}_t)). \tag{19}$$

We first prove that

$$\varepsilon_t^F = \varepsilon_t^H - \mathcal{I}_{p_{data}}(\mathbf{x}_t; \mathbf{x}_T|\mathbf{x}_{t-1}), \tag{20}$$

where $\mathcal{I}_{p_{data}}(\mathbf{x}_t; \mathbf{x}_T|\mathbf{x}_{t-1})$ is the conditional mutual information under q.

Since $\mathbf{x}_{t-1}$ and $\mathbf{q}$ can be considered as fixed variables at the step t, we omit them in the notation for brevity. According to the definition of KL divergence, we have:

$$\varepsilon_t^F \triangleq D_{KL}\big(p_{data}(\mathbf{x}_t), \pi_F(\mathbf{x}_t)\big) = \sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[\log p_{data}(\mathbf{x}_t) - \log \pi_F(\mathbf{x}_t)\Big]. \tag{21}$$

By construction

$$\pi_F(\mathbf{x}_t) = \frac{\exp(S(\mathbf{x}_t))}{Z_t}, \qquad S(\mathbf{x}_t) = C_{\text{local}}(\mathbf{x}_t) + C_{\text{global}}(\mathbf{x}_t), \qquad Z_t = \sum_{\mathbf{x}_t'} \exp(S(\mathbf{x}_t')). \tag{22}$$

Hence

$$\log \pi_F(\mathbf{x}_t) = S(\mathbf{x}_t) - \log Z_t = C_{\text{local}}(\mathbf{x}_t) + C_{\text{global}}(\mathbf{x}_t) - \log Z_t. \tag{23}$$

By inserting Eq. 23 into 21 and split the sum,

$$\begin{aligned}\varepsilon_t^F &= \sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[\log p_{data}(\mathbf{x}_t) - C_{\text{local}}(\mathbf{x}_t) - C_{\text{global}}(\mathbf{x}_t) + \log Z_t\Big] \\ &= \underbrace{\sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[\log p_{data}(\mathbf{x}_t) - C_{\text{local}}(\mathbf{x}_t)\Big]}_{\text{Term A}} - \underbrace{\sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[C_{\text{global}}(\mathbf{x}_t) - \log Z_t\Big]}_{\text{Term B}}.\end{aligned} \tag{24}$$

By definition $C_{\text{local}}(\mathbf{x}_t) = \log p_\theta(\mathbf{x}_t)$, so

$$\text{Term A} = \sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[\log p_{data}(\mathbf{x}_t) - \log p_\theta(\mathbf{x}_t)\Big] = D_{KL}\big(p_{data}(\mathbf{x}_t), p_\theta(\mathbf{x}_t)\big) \triangleq \varepsilon_t^H. \tag{25}$$

First recall

$$C_{\text{global}}(\mathbf{x}_t) = \mathbb{E}_{x'\sim p_\theta} \log p_\theta(x' \mid \mathbf{x}_t) = \sum_{\mathbf{x}_T} p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) \log p_\theta(\mathbf{x}_T \mid \mathbf{x}_t). \tag{26}$$

Therefore

$$\begin{aligned}\text{Term B} &= \sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[C_{\text{global}}(\mathbf{x}_t) - \log Z_t\Big] \\ &= \sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t)\Big[\sum_{\mathbf{x}_T} p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) \log p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) - \log Z_t\Big].\end{aligned} \tag{27}$$

Next we use the identity $Z_t = p_\theta(\mathbf{x}_T)$ (marginal over $\mathbf{x}_t'$), which follows from

$$Z_t = \sum_{\mathbf{x}_t'} \exp(S(\mathbf{x}_t')) = \sum_{\mathbf{x}_t'} p_\theta(\mathbf{x}_t', \mathbf{x}_T) = p_\theta(\mathbf{x}_T). \tag{28}$$

Hence

$$\log Z_t = \log p_\theta(\mathbf{x}_T) = \sum_{\mathbf{x}_T} p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) \log p_\theta(\mathbf{x}_T), \tag{29}$$

where the second equality holds because $p_\theta(\mathbf{x}_T)$ does not depend on $\mathbf{x}_t$ and $\sum_{\mathbf{x}_T} p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) = 1$. Inserting this into Eq. 27 gives

$$\begin{aligned}
\text{Term B} &= \sum_{\mathbf{x}_t} p_{data}(\mathbf{x}_t) \sum_{\mathbf{x}_T} p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) \Big[\log p_\theta(\mathbf{x}_T \mid \mathbf{x}_t) - \log p_\theta(\mathbf{x}_T)\Big] \\
&= \sum_{\mathbf{x}_t, \mathbf{x}_T} p_{data}(\mathbf{x}_t, \mathbf{x}_T) \Big[\log \frac{p_\theta(\mathbf{x}_T \mid \mathbf{x}_t)}{p_\theta(\mathbf{x}_T)}\Big] \\
&= \sum_{\mathbf{x}_t, \mathbf{x}_T} p_{data}(\mathbf{x}_t, \mathbf{x}_T) \Big[\log \frac{p_{data}(\mathbf{x}_t, \mathbf{x}_T)}{p_{data}(\mathbf{x}_t) p_{data}(\mathbf{x}_T)}\Big] \qquad \big(\text{replace } p_\theta \text{ with } p_{data} \text{ inside log}\big) \\
&= \mathcal{I}_{p_{data}}(\mathbf{x}_t; \mathbf{x}_T). \qquad (30)
\end{aligned}$$

Inserting Eq. 25 and 30 into Eq. 24 yields

$$\varepsilon_t^F = \varepsilon_t^H - \mathcal{I}_{p_{data}}(\mathbf{x}_t; \mathbf{x}_T), \qquad (31)$$

which completes the proof of Eq. 20. Using the chain rule for KL divergence over sequences we have

$$D_{KL}\big(p_{data}(\mathbf{x}), p_{\pi_F}\big) = \sum_{t=1}^{T} \mathbb{E}_{p_{data}(\mathbf{x}_{t-1})}\big[\varepsilon_t^F\big] = \sum_{t=1}^{T} \mathbb{E}_{p_{data}(\mathbf{x}_{t-1})}\big[\varepsilon_t^H - \mathcal{I}_{p_{data}}(\mathbf{x}_t; \mathbf{x}_T | \mathbf{x}_{t-1})\big]. \qquad (32)$$

Recognising the definitions

$$\sum_{t=1}^{T} \mathbb{E}_{p_{data}(\mathbf{x}_{t-1})} \varepsilon_t^H = D_{KL}\big(p_{data}(\mathbf{x}), p_{\pi_H}\big), \quad \sum_{t=1}^{T} \mathbb{E}_{p_{data}(\mathbf{x}_{t-1})} \mathcal{I}_{p_{data}}(\mathbf{x}_t; \mathbf{x}_T | \mathbf{x}_{t-1}) = \Delta_{\text{total}}, \qquad (33)$$

we obtain

$$D_{KL}\big(p_{data}(\mathbf{x}), p_{\pi_F}\big) = D_{KL}\big(p_{data}(\mathbf{x}), p_{\pi_H}\big) - \Delta_{\text{total}}, \qquad (34)$$

which finishes the proof of Theorem 1. □

C Details of FDM

Algorithm 1 Foreseeing Decoding Method (FDM)

Input: User query $\mathbf{q}$, the partially decoded sequence $\mathbf{x}_{t-1}$, pruning threshold γ, width K, well-trained models θ, the number of tokens for decoding n.
// Get the candidate tokens of the undecoded positions.
`Candidate` = $[\mathbf{1}[\mathbf{x}_{t-1} = \texttt{Mask}] \odot \arg\max p_\theta(\mathbf{q}, \mathbf{x}_{t-1})]$
for x_t in `Candidate` **do**
  if $p_\theta(\mathbf{q}, \mathbf{x}_{t-1})[x_t] \leq \gamma$ **then**
    Delete x_t from `Candidate`.
  end if
end for
Priority Queue $\{\mathbf{x}_t\}$ of the priority C_{local} by foreseeing decoding n tokens from `Candidate`.
// Narrow the search space with the width K.
Λ = Top-K($\{\mathbf{x}_t\}$)
if $\Lambda = \varnothing$ **then**
  $\mathbf{x}_t = \arg\max C_{local}(\mathbf{x}_t)$
else
  Calculate $C_{global}(\mathbf{x}_t)$ for each $\mathbf{x}_t$ in Λ
  $\mathbf{x}_t = \arg\max_{\mathbf{x}_t \in \Lambda} \{C_{local}(\mathbf{x}_t) + C_{global}(\mathbf{x}_t)\}$
end if
Output: The decoded answer $\mathbf{x}_t$ at step t

D DETAILS OF FDM-A

Algorithm 2 Acceleration with FDM (FDM-A)

Input: User query $\mathbf{q}$, fully masked sequence $\mathbf{x}_0$, pruning threshold γ_1, searching width K_1, stage division coefficients η_1 and η_2, upper bound for the decoding number N, well-trained models θ.
Initialize $\texttt{FDM}$ with $K = K_1$ and get $\texttt{FDM}_1$.
Initialize $\texttt{FDM}$ with $K = 1$ and get $\texttt{FDM}_2$.
$t = 1$
while $\texttt{Mask}$ not in $\mathbf{x}_t$ **do**
  if $\texttt{NUM}\big([\mathbf{x}_{t-1} = \texttt{Mask}]p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1\big) = 0$ **then**
    $\mathbf{x}_t = \texttt{FDM}_1(\mathbf{x}_{t-1}, \mathbf{q}, n = 1, \gamma = \gamma_1)$
  else if $\texttt{NUM}\big([\mathbf{x}_{t-1} = \texttt{Mask}]p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1\big) \geq N$ **then**
    $\mathbf{x}_t = \texttt{FDM}_2(\mathbf{x}_{t-1}, \mathbf{q}, n = N, \gamma = 1.0)$
  else if $\texttt{NUM}\big(\eta_2 < [\mathbf{x}_{t-1} = \texttt{Mask}]p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) \leq \eta_1\big) = 0$ **then**
    $\mathbf{x}_t = \texttt{FDM}_2(\mathbf{x}_{t-1}, \mathbf{q}, n = \texttt{NUM}([\mathbf{x}_{t-1} = \texttt{Mask}]p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1), \gamma = 1.0)$
  else
    $\mathbf{x}_t = \texttt{FDM}_1(\mathbf{x}_{t-1}, \mathbf{q}, n = \texttt{NUM}([\mathbf{x}_{t-1} = \texttt{Mask}]p_\theta(\mathbf{q}, \mathbf{x}_{t-1}) > \eta_1), \gamma = \eta_2)$
  end if
  $t = t + 1$
end while
Output: Generated answer $\mathbf{x}_t$.

E SYSTEM PROMPTS FOR EVALUATION

GSM8K

You are a math expert. You will be given a question to solve. Solve it step by step. Wrap the final answer in a \\boxed{}.
Respond in the following format:
<reasoning >
Your reasoning here
</reasoning >
<answer>
\\boxed{...}
</answer>

ARC

You will be given a multiple choice question and a list of options. You need to first reason step by step, and then select the correct option (A, B, C, D). Wrap the single letter of the correct option in \\boxed{...}. Respond in the following format:
<reasoning>
Your reasoning here
</reasoning>
<answer>
\\boxed{...}
</answer>

Countdown

Using only the provided numbers, create an arithmetic expression that evaluates to exactly the provided target number. You may use the operations +, -, *, and / as needed, but each number must be used exactly once. Think step-by-step. After reasoning, provide only your final expression inside \\boxed{}
tags without including an equals sign or the target number. For example: \\boxed{a + b * c}
Respond in the following format:
<reasoning>
Your reasoning here
</reasoning>
<answer>
\\boxed{...}
</answer>

HumanEval

Write a solution to the following problem and make sure that it passes the tests:
python

F THE EXAMPLE CASE

Question in the ARC benchmark

Which of the following is the primary way that evergreen trees planted on the north and west sides of residences in Massachusetts help reduce winter heating costs?
A They insulate against cold air
B They optimize solar heating.
C They reduce radiant cooling.
D They act as a windbreak.

Decoding with the highest probability (Incorrect)

Evergreen trees planted on the north and west sides of residences in Massachusetts help reduce winter heating costs primarily by acting as a physical barrier against cold cold air. This trees' asymmetrical and dense foliage provide a a natural insulation, preventing cold air from entering the home, thereby reducing the need for heating heating.
</reasoning>
<answer>
\boxed{A}
</answer>

Decoding with the highest margin (Incorrect)

Evergreen trees planted on the north and west sides of residences in Massachusetts can help reduce winter heating costs by providing as much solar heat as possible. This is because the trees can capture and store sunlight, which can be used to heat the home during the winter months. By maximizing the amount of solar heat, the trees can help reduce the need for additional heating systems.
</reasoning>

<answer>
B

Decoding with the lowest entropy (Incorrect)

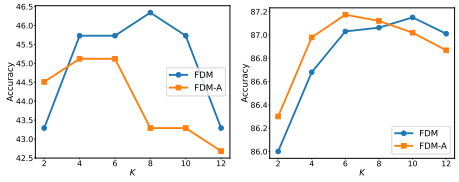
Evergreen trees are planted on the north and west sides of residences in Massachusetts to help reduce winter heating costs. The primary reason for this benefit is due to the insulation provided by the trees. The trees provide insulation against the cold, reducing the amount of heat escaping the house and keeping the interior temperature slightly warmer. This insulation reduces the need for heating, thereby making the home more efficient.
<answer>
\boxed{A}

Decoding with FDM (Correct)

Evergreen trees planted on the north and west sides of residences in Massachusetts can help reduce winter heating costs by acting as a windbreak. These trees can block cold air from entering the home, reducing the need for heating. Additionally, evergreen trees can also provide insulation against cold air, but the primary benefit is their role as a windbreak.
</reasoning>
<answer>
\boxed{D}
</answer>

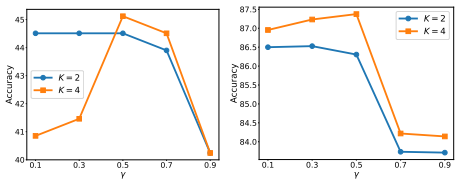
G Ablation Studies on the HumanEval and ARC Benchmarks

We perform experiments on the HumanEval and ARC benchmarks with the experimental settings in Section 5.2. We observe that the optimal hyperparameter has good transferability across datasets.



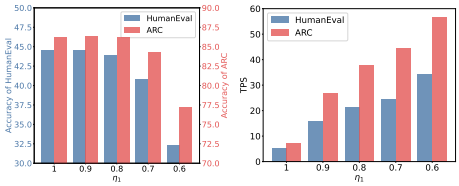
(a) HumanEval (b) ARC

Figure 8: The influence of K on model performance on HumanEval and ARC benchmarks.



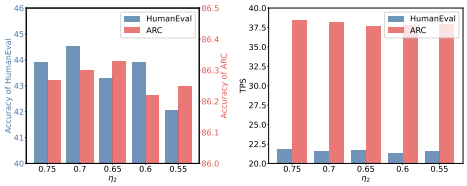
(a) HumanEval (b) ARC

Figure 9: The influence of γ on model performance on HumanEval and ARC benchmarks.



(a) Accuracy (b) TPS

Figure 10: The influence of η_1 on Accuracy and TPS on HumanEval and ARC benchmarks.



(a) Accuracy (b) TPS

Figure 11: The influence of η_2 on Accuracy and TPS on HumanEval and ARC benchmarks.

H Theoretical Analysis for Performance Degradation with a Extremely Large K

We next present a theoretical analysis to support that the degradation in performance is due to the consequence of noise-amplified selection in the mismatch between the model and natural distributions. At each decoding step, we have K candidate tokens whose true future returns are

$$s_1, \dots, s_K \in \mathbb{R}, \quad s_* = \max_i s_i, \text{where } s_i \triangleq \log p_{\text{data}}(\mathbf{x}_t = \text{i-th candidate token} \mid \mathbf{x}_{t-1}, \mathbf{q}). \quad (35)$$

When K is large enough, we can assume that the token with the highest possibility in the real data distribution is already in the candidates. The model in our algorithm only observes confidence scores, which are noisy estimates of true future returns:

$$\hat{s}_i = s_i + \xi_i \quad (36)$$

Here we hypothesize that ξ_i is independent and identically distributed (i.i.d.) Gaussian random variables:

$$\xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad (37)$$

and $\hat{s}_i$ is equal to the sum of the local and global confidence of the i-th candidate token:

$$\hat{s}_i = C_{local}(\text{i-th candidate token}) + C_{global}(\text{i-th candidate token}) \quad (38)$$

According to Eq. 8, our algorithm selects the candidate with the highest $\hat{s}_i$:

$$j = \underset{i \in [K]}{argmax}\, \hat{s}_i. \quad (39)$$

For any sub-optimal candidate i with confidence gap $\Delta_i = s_* - s_i > 0$, the probability that noise flips the ranking is

$$\Pr\big(\hat{s}_i > \hat{s}_*\big) = \Pr\big(\xi_i - \xi_* > \Delta_i\big) = \Phi\Big(-\frac{\Delta_i}{\sqrt{2}\sigma}\Big), \quad (40)$$

where $\Phi(\cdot)$ is the standard normal tail CDF(Cumulative Distribution Function). Union-bounding over the $K-1$ sub-optimal candidates yields

$$\Pr(\text{select non-optimal}) \le \sum_{i \neq *} \Phi\Big(-\frac{\Delta_i}{\sqrt{2}\sigma}\Big). \quad (41)$$

The RHS(right hand side) increases monotonically with K whenever there exist $\Delta_i = O(\sigma)$. On the other hand, let $\xi_{(K)} = \max_{i \le K} \xi_i$. The Extreme-value Theory gives

$$\mathbb{E}[\xi_{(K)}] \approx \sigma\sqrt{2 \ln K}. \quad (42)$$

Hence, the selected estimate satisfies

$$\mathbb{E}[\hat{s}_j] \approx s_* + \sigma\sqrt{2 \ln K}, \quad (43)$$

while the true return of the selected candidate satisfies:

$$\mathbb{E}[s_j] \le s_* + \underbrace{\mathbb{E}[\xi_{(K)} - \xi_j]}_{\ge 0}. \quad (44)$$

Consequently, the expected regret is:

$$\mathbb{E}[\Delta_j] = \mathbb{E}[s_* - s_j] \ge C\sigma\sqrt{\ln K} \quad (45)$$

which grows with K, providing a quantitative lower bound for the "winner's curse" in the decoding process. Because each incorrect selection changes the subsequent context, the regret accumulates along the generation process. Early mistakes (where K is too large in our schedule) therefore have a negative effect, explaining the empirical observation in Fig. 4 and 8.