Inference Strategies for Reasoning of Discrete Diffusion Language Models: An Empirical Analysis

Jin Woo Koo, Jungwoo Lee

Seoul National University

jinukoo@snu.ac.kr, junglee@snu.ac.kr

이산 확산 언어 모델 추론을 위한 디코딩 전략: 실증적 분석

구진우, 이정우
서울대학교

Abstract

Discrete diffusion language models have shown that with on par performance with autoregressive language models, they have strengths of parallel generation and editing wrong generated tokens. In this research, we compared different types of inferencing techniques to find the best inference strategy. With experimenting with two open-sourced discrete diffusion language models, LLaDA 8B model and Dream 7B model, we found that block diffusion style of inferencing achieves the best performance on GSM8K, math reasoning benchmark.

Ⅰ. Introduction

Autoregressive transformer based language models have shown great success in language modeling.[1] However, they have limitations due to their left to right nature: they suffer from sequential inference latency as they have to generate tokens sequentially, and they do not have the ability to correct wrong generated tokens.

Discrete diffusion language models are a new paradigm of language modeling that adapted discrete diffusion to language modeling. Combined with masked language modeling, discrete diffusion language models show on par performance with autoregression language models.[7] However, unlike autoregressive language models, they can generate tokens in parallel, look over past and future to generate tokens, and edit wrong tokens after generation of that token. As they have advantages that autoregressive language models lack, they still need more research.

In this work, we try to answer the following research question: which method is best for inferencing discrete diffusion language models? With experiment, we found that block diffusion inference strategy is the best strategy to use for math reasoning.

Ⅱ. Method

We experimented with two discrete diffusion language models: LLaDA 8B model and Dream 7B model.[2,3]. We evaluated them on GSM8K Math reasoning benchmark.[4]

We experimented three types of inference strategies: pure diffusion, block diffusion, and autoregressive generation. Pure diffusion inference strategy randomly selects tokens at every step. Block diffusion strategy divides the whole sequence into blocks and performs diffusion in that block. So, at every step, block diffusion only predicts tokens in the current block whereas pure diffusion can predict any token in the sequence.[6] Lastly, autoregressive generation generates similarly to decoder-only language models; one token at a time, sequentially.

We used fast-dLLM github repository for the experiments.[5] The results are in Table 1.

Table 1. Experiment Results

	LLaDA Base	LLaDA Instruct	Dream Base	Dream Instruct	Average
Pure Diffusion	2.05	23.35	63.31	78.09	41.7
Block Diffusion	35.18	73.31	64.29	78.54	62.83
Autoregressive	25.93	75.06	63.31	78.09	60.5975

We can see from the result that block diffusion inference strategy showed the best performance. By generating tokens near already-produced content, it better leverages contextual information to produce correct answers.

Ⅲ. Conclusion

In this work, we experimented to find out the best method for inferencing discrete diffusion language models. By experiment, we found that block diffusion generation technique can achieve best performance for math reasoning benchmark across different discrete diffusion language models.

ACKNOWLEDGMENT

This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2025-02305453(15%), RS-2025-02273157(15%), RS-2025-25442149(15%) RS-2021-II211343(15%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2026.

REFERENCES

[1] Brown, B. Tom, et al. "Language Models are Few-Shot Learners," Advances in neural information processing systems 33 (2020), pp. 1877- 1901, 2020.

[2] Nie, N. et al. " Large Language Diffusion Models", 2025, (https://arxiv.org/abs/2502.09992).

[3] Ye, J. et al. "Dream 7B: Diffusion Large Language Models", 2025, (https://arxiv.org/abs/2508.15487).

[4] Cobbe, K. et al. "Training Verifiers to Solve Math Word Problems", 2021, (https://arxiv.org/abs/2110.14168).

[5] Wu, C. et al. "Fast-dLLM: Training-free Acceleration of Diffusion LLM by Enabling KV Cache and Parallel Decoding", 2025, (https://arxiv.org/abs/2505.22618).

[6] Arriola, M. et al. "Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models", 2025, (https://arxiv.org/abs/2503.09573).

[7] Sahoo, S. et al. "Simple and Effective Masked Diffusion Language Models", 2024, (https://arxiv.org/abs/2406.07524)