Empirical Analysis of Decoding Biases in Masked Diffusion Models

Pengcheng Huang[1], Tianming Liu[1], Zhenghao Liu[1], Yukun Yan[2], Shuo Wang[2], Tong Xiao[1], Zulong Chen[3], Maosong Sun[2]

1 Northeastern University, China 2 Tsinghua University-NLP Lab 3 Alibaba Group

ACL 2026 SAN DIEGO JULY 2-7

> 7%
avg. gain over the best decoding baseline

44.7 ≈ 45.3
LLaDA-1.5 + UNCODE vs. Qwen-2.5-7B (AR)

3 × 7
MDM backbones × reasoning & planning tasks

0
extra training — plug-and-play, decoding-only

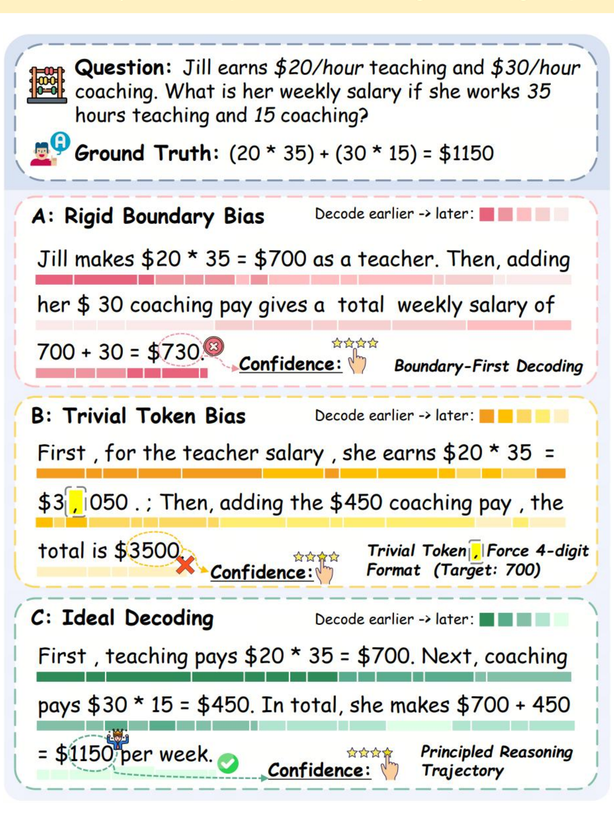
Motivation and Background

- **Masked Diffusion Models (MDMs)** decode by iterative unmasking — any-order, multi-token, non-autoregressive.
- The **unmasking order is decisive**: standard MDMs greedily unmask the least-uncertain positions first.

The Two Decoding Biases

(1) **Rigid Boundary Bias:** boundary tokens are always decoded first, collapsing inward — the answer is fixed before the rationale exists.

(2) **Trivial Token Bias:** high-frequency, low-information tokens (punctuation, fillers) are over-prioritized, wasting decoding steps.



Uncertainty-based decoding prioritizes (A) boundaries and (B) trivial tokens, deviating from the (C) ideal reason-then-answer path.

UNCODE: Unmasking Calibration

A lightweight, **training-free** calibration of the unmasking priority s_t^i via two complementary priors:

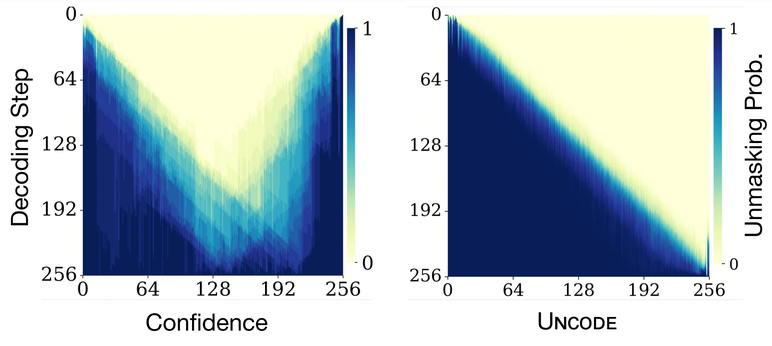
$$\tilde{s}_t^i = \underbrace{\mathcal{P}^i}_{\text{position}} \cdot \underbrace{\mathcal{S}_t^i}_{\text{semantics}} \cdot \underbrace{\mathcal{F}\big(p_\theta(\cdot \mid x_t, i)\big)}_{\text{raw uncertainty}}$$

Positional Trajectory Prior: $P^i = e^{-\lambda i}$ — discourages boundary-first decoding; λ interpolates between flexible any-order and left-to-right generation.

Semantic Informativeness Prior: $S_t^i = \min(-\log_{P_D}(\overline{x}_t^i), \alpha)$, down-weights frequent trivial tokens via corpus self-information (clipped at α).

⇒ Keeps parallel any-order decoding while **globally reshaping the trajectory** and **promoting informative content**.

Trajectory Reshaping



Experimental Setup

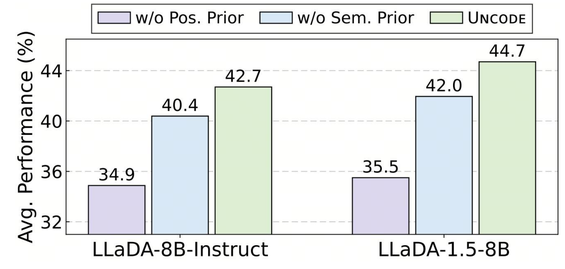
- **3 MDM backbones**: LLaDA-8B-Instruct, LLaDA-1.5-8B, Dream-7B.
- **7 benchmarks**: HumanEval, MBPP (code); GSM8K, MATH500 (math); GPQA (science); Countdown, Sudoku (planning).
- **8 decoding baselines**: Uniform, Confidence, Entropy, Margin, EB-Sampler, Semi-AR, Fast-dLLM.

Experimental Results

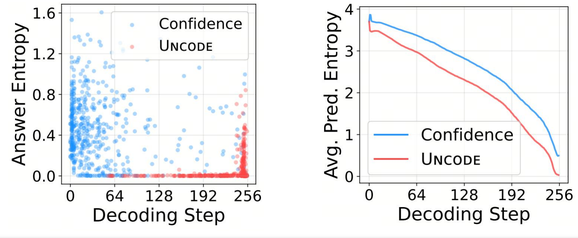
- **Overall Performance**
  - UNCODE outperforms all baseline models across datasets.

Method	GSM8K	MATH	Sudoku	C.down	Avg.
Autoregressive reference					
Qwen-2.5-7B	71.9	**64.2**	0.0	7.7	45.3
LLaDA-8B-Instruct					
Semi-AR (best baseline)	77.9	27.6	0.0	32.6	35.7
UNCODE	**79.2**	34.8	**29.8**	**36.3**	**42.7**
LLaDA-1.5-8B					
Semi-AR (best baseline)	80.7	34.2	0.0	32.4	37.1
UNCODE	**82.2**	37.4	**33.4**	35.0	**44.7**

- **Ablation and Analysis**
  - Both priors matter



  - Defers answer tokens to later steps with higher confidence, and reduces global uncertainty faster than the baseline.



Resources

Paper Code Email Wechat

Explore more interesting experiments in our paper and code, and feel free to contact us at pengcheng.neu@outlook.com or via WeChat