Beyond Confidence: Adaptive and Coherent Decoding for Diffusion Language Models

Kecheng Chen, Ziru Liu, et al.

City University of Hong Kong & Huawei Research

December 12, 2025

引言与动机

扩散语言模型(DLM) 具备全局感知规划与并行推理等潜力。然而，现有的采样过程面临关键局限性：

- **易陷入局部最优：** 现有的单步指标（如置信度或熵）缺乏前瞻性，容易导致模型在当前步骤选择局部最优解。
- **缺乏理论基础：** 当前的采样过程缺乏与采样错误率的直接理论联系，难以保证解码质量。
- **解码预算固化：** 统一的解码预算效率低下，无法适应简单生成与复杂的逻辑推理切换的需求。

后果： 导致采样轨迹上下文不一致，生成效率低下。

连贯上下文解码(CCD)

为了解决上述挑战，我们提出了**CCD** 推理框架。

核心创新点

1. **轨迹矫正机制：** 利用历史上下文增强序列连贯性，从而提前拒绝次优的路径。
2. **理论基础：** 通过上下文与Token预测之间的**条件互信息**来对一致性进行建模。
3. **自适应采样(CCD-DS)：** 根据一致性度量动态调整每步的解码预算（Unmasking Budget），显著加速推理。

预备知识：基础采样过程

扩散语言模型(DLM) 通过迭代去噪，从完全掩码状态$\mathbf{x}_T$ 逐步生成干净数据$\mathbf{x}_0$。

目前的解码方案通常采用统一的预算$b_t \approx N/T$。在每一步t，模型根据**单步预测分布**的置信度选择b_t 个Token 进行去掩码（解码）。

标准采样公式

采样过程形式化为：

$$x_{t,i} = \begin{cases} \arg\max_{k \in \mathbb{X}} \ p_{t,i}^k & \text{若} i \in \mathcal{J}_t \\ x_{t+1,i} & \text{其他情况} \end{cases}$$

其中解码集合$\mathcal{J}_t$ 由最大化确定性（负熵）决定：

$$\mathcal{J}_t = \arg\max_{\mathcal{S} \subset \mathcal{K}_t, |\mathcal{S}| = b_t} \sum_{i \in \mathcal{S}} -H(p_{t,i})$$

注：$H(p_{t,i})$ 为香农熵。现有方法主要依赖当前的单步预测，容易受局部误差影响。

理论洞察：互信息

现有方法使用单步预测分布$\hat{p}_\theta(x_i|\mathbf{c}_{\cdot,i}, \mathbf{s})$ 近似目标分布。

我们的洞察： 通过对解码轨迹上的上下文进行积分，近似真实的边缘分布。

$$p(x_i \mid \mathbf{s}) \approx \overline{p}(x_i \mid \mathbf{s}) \triangleq \frac{1}{T-t+1} \sum_{k=0}^{T-t} \hat{p}_\theta(x_i \mid \mathbf{c}_{T-k,i}, \mathbf{s}) \tag{1}$$

这在理论上关联到了**条件互信息**：

$$H(x_i|\mathbf{s}) = H(x_i|\mathbf{c}, \mathbf{s}) + I(x_i; \mathbf{c}|\mathbf{s})$$

这意味着我们应当优先选择那些置信度高，且在不同解码步骤间预测保持一致的*Token*。

通过CCD控制误差界

我们证明了使用CCD可以有效地控制采样误差的上界。

Proposition 2 (Sampling Error Bound)

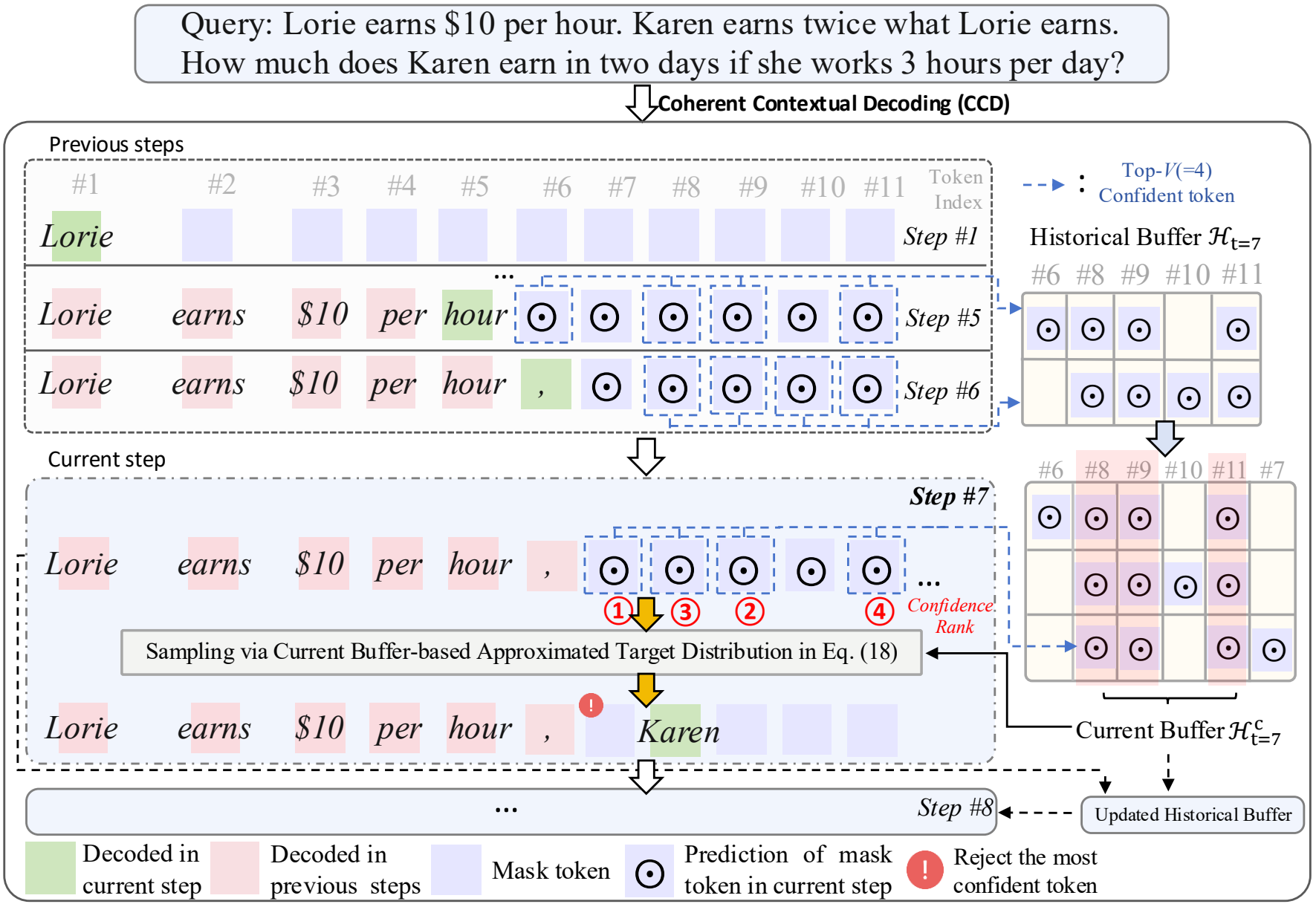
在条件生成下，随着$t \to 0$，采样误差界受以下公式控制：

$$\mathbb{E}\left[\mathrm{KL}(p(\bar{\mathbf{x}}|\mathbf{s}) \,\|\, p(\hat{\mathbf{x}}|\mathbf{s}))\right] \leq \frac{G}{T} \sum_{i=1}^{N} \underbrace{\left[\frac{1}{T-t+1} \sum_{k=0}^{T-t} I(\bar{x}_i; \mathbf{c}_{T-k,i}|\mathbf{s}) \right]}_{\text{解码轨迹上的平均互信息}} + \varepsilon_{\text{train}}$$

理论意义：

- 该界表明，采样误差取决于Token 与其不断演变的上下文之间的平均互信息**(Mutual Information)**。
- 我们的方法通过在解码轨迹上平均互信息，有效地最小化了这一误差界，从而在理论上保证了生成的连贯性。

CCD 框架概览


Query: Lorie earns $10 per hour. Karen earns twice what Lorie earns. How much does Karen earn in two days if she works 3 hours per day?
Coherent Contextual Decoding (CCD)
Previous steps
Token Index
Step #1
Step #5
Step #6
Top-V(=4) Confident token
Historical Buffer $\mathcal{H}_{t=7}$
Current step
Step #7
Confidence Rank
Sampling via Current Buffer-based Approximated Target Distribution in Eq. (18)
Current Buffer $\mathcal{H}^c_{t=7}$
Updated Historical Buffer
Step #8
Decoded in current step
Decoded in previous steps
Mask token
Prediction of mask token in current step
Reject the most confident token

实现：滑动窗口历史缓冲区

直接存储所有分布会导致内存开销过大。

高效解决方案

- **历史缓冲区($\mathcal{H}_t$)：** 仅存储最近d 次迭代中Top-V 的高置信度Token。
- **过滤噪声：** 自动过滤掉早期扩散步骤中不稳定的无效预测。
- **一致性检查：** 只有在缓冲区内跨步持续出现在Top-V 集合中的Token 才会被选为解码候选。

内存复杂度降低： 从$O((T-t) \times N \times |\mathbb{X}|)$ 降至$O(d \times V \times |\mathbb{X}|)$。

自适应采样预算(CCD-DS)

观察：

- 标准DLM 使用统一预算（图中虚线）。
- 解码过程存在“停滞期”（Plateaus），期间仅生成EOS Token，效率极低。

CCD-DS 策略：

- 根据一致性Token 的数量动态变化解码步长。
- 上下文不敏感区域（如模板）分配大预算；复杂推理分配小预算。

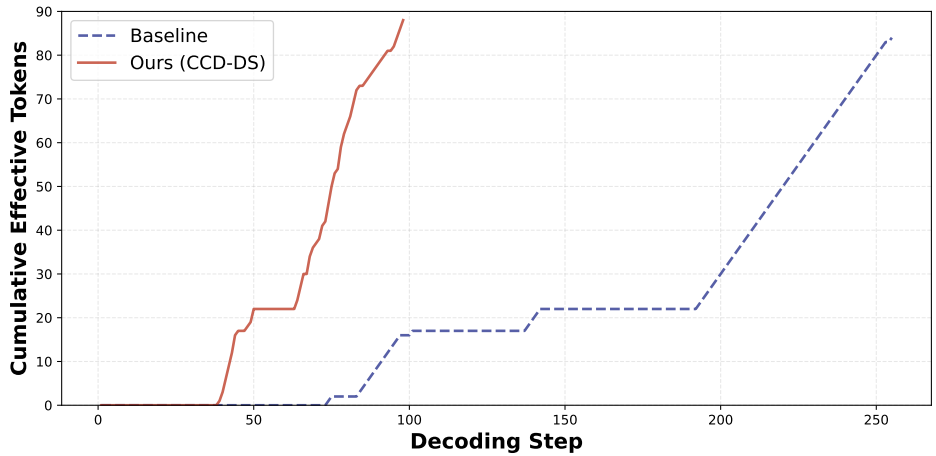


Figure: 累积有效Token 数。注意基线方法的停滞期vs CCD-DS 的持续增长。

采样预算分析

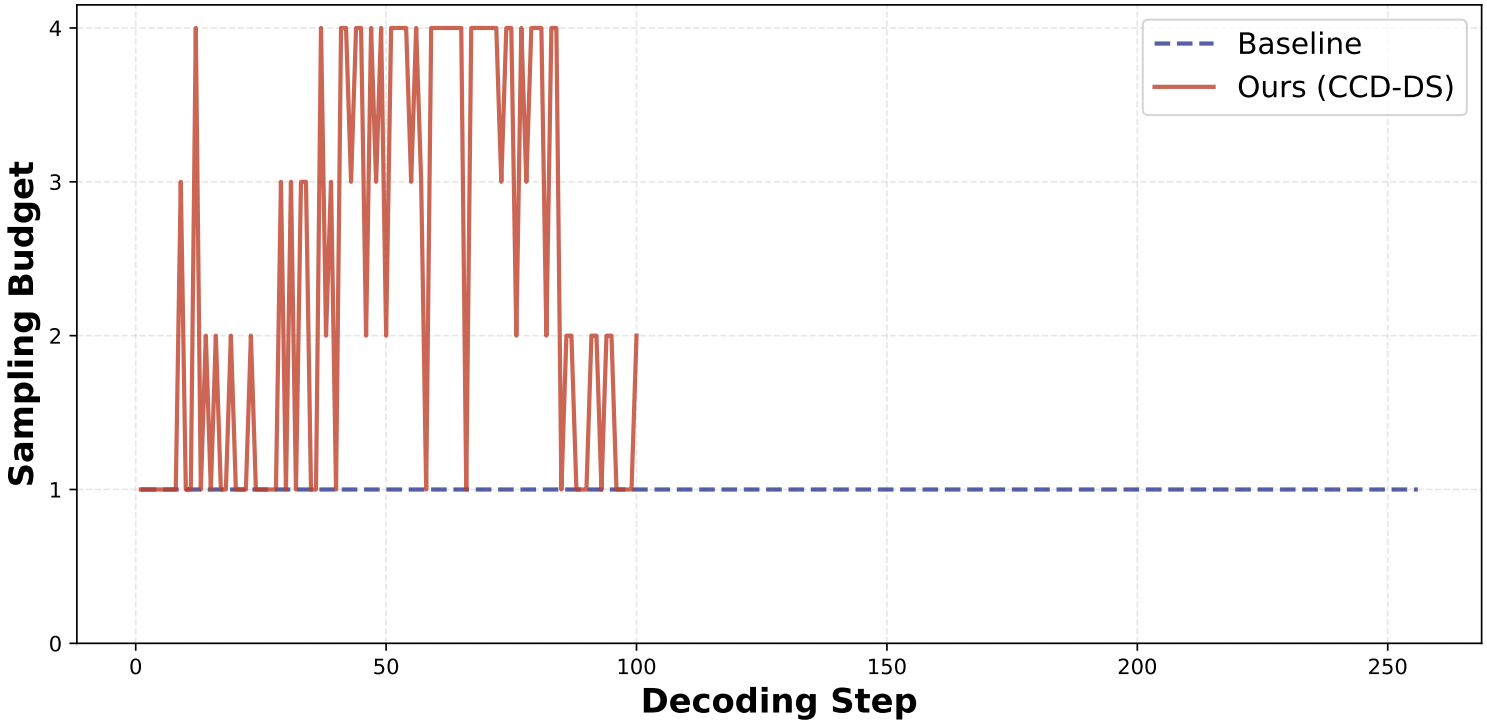


Figure: 采样预算对比。基线为固定预算($b_t = 1$)。CCD-DS 根据生成难度动态调整预算（最高为4），从而加速解码过程。

实验设置

- **基础模型:**
  - LLaDA-8B-Instruct
  - Dream-7B-Instruct
- **基准测试(Benchmarks):**
  - 数学推理： GSM8K, MATH
  - 代码生成： HumanEval, MBPP
  - 规划： Trip Plan
- **基线对比:**
  - 采用统一预算的标准采样过程。
  - Dream 系列使用负熵，LLaDA 系列使用最大概率作为置信度。

主要结果

Task	Method	Inference Efficiency		Performance	Method	Inference Efficiency		Performance
		Diffusion steps↓	Gains↑	Score↑		Diffusion steps↓	Gains↑	Score↑
Mathematics Reasoning								
GSM8K	LLaDA Instruct	512	1.00×	74.30	Dream Instruct	256	1.00×	81.01
	+ CCD	512	1.00×	**75.30** +1.00	+ CCD	256	1.00×	82.26 +1.25
	+ CCD-DS	393.0 −119.0	1.31× +0.31	75.22 +0.92	+ CCD-DS	141.2 −114.8	1.82× +0.82	**82.51** +1.50
Math	LLaDA Instruct	512	1.00×	37.00	Dream Instruct	512	1.00×	40.90
	+ CCD	512	1.00×	**37.20** +0.20	+ CCD	512	1.00×	**41.20** +0.30
	+ CCD-DS	378.2 −133.8	1.35× +0.35	**37.20** +0.20	+ CCD-DS	340.2 −171.8	1.58× +0.58	**41.20** +0.30
Code Generation								
HumanEval	LLaDA Instruct	512	1.00×	36.50	Dream Instruct	768	1.00×	52.66
	+ CCD	512	1.00×	**38.41** +1.91	+ CCD	768	1.00×	**57.31** +4.65
	+ CCD-DS	332.0 −180.0	1.54× +0.54	38.40 +1.90	+ CCD-DS	253.2 −514.8	3.04× +2.04	56.71 +4.05
MBPP	LLaDA Instruct	256	1.00×	39.20	Dream Instruct	1024	1.00×	58.00
	+ CCD	256	1.00×	39.20	+ CCD	1024	1.00×	58.00
	+ CCD-DS	211.20 −44.8	1.24× +0.24	**39.20** +0.00	+ CCD-DS	270.20 −753.80	3.78× +2.78	**58.00** +0.00
Planing								
Trip Plan	LLaDA Instruct	256	1.00×	10.40	Dream Instruct	256	1.00×	15.10
	+ CCD	256	1.00×	10.80 +0.40	+ CCD	256	1.00×	16.93 +1.83
	+ CCD-DS	112.5 −143.5	2.27× +1.27	**11.50** +1.10	+ CCD-DS	75.20 −180.20	3.48× +2.48	**19.01** +3.91

Figure: LLaDA和Dream在5个不同Benchmarks上的表现

超参数分析

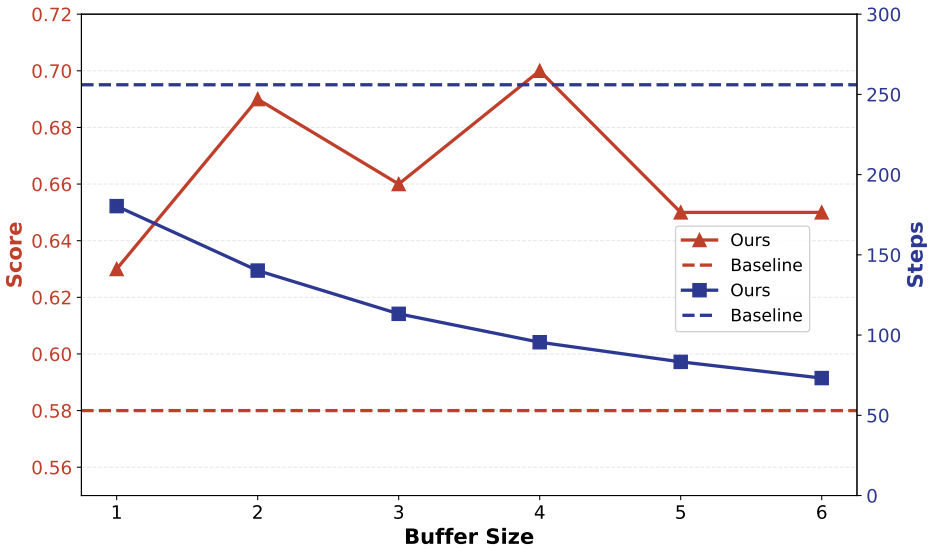


Figure: 缓冲区大小的影响。Size=4 时达到最佳平衡。

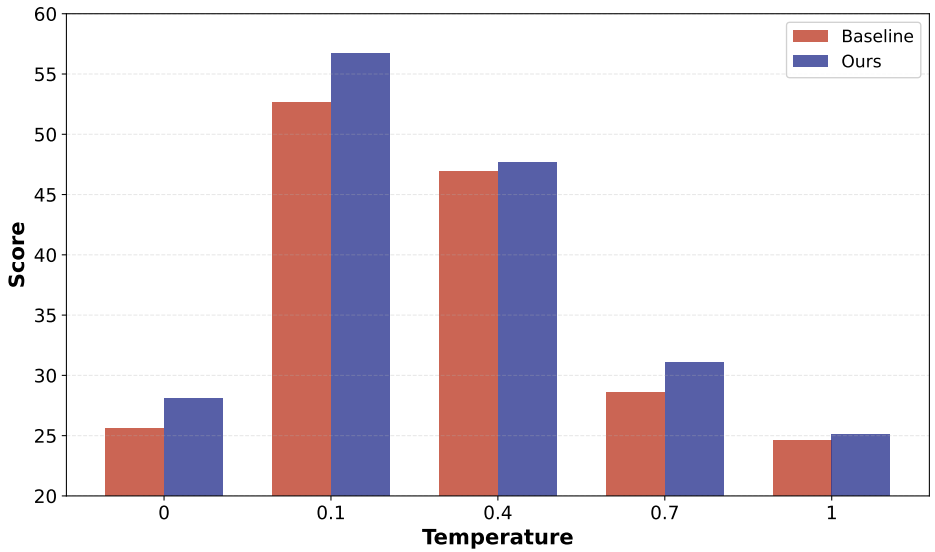


Figure: 温度鲁棒性分析。CCD 在不同温度下均优于基线。

案例分析：为何CCD 更有效?

Task: *Lorie earns $10 per hour. Karen earns twice what Lorie earns. How much does Karen earn in two days if she works 3 hours per day?* `Incorrect` `Correct` ⊗ Mask token

Baseline: Lorie earns $10 per hour, so in 3 hours she earns $10 x 3 = $30.
Karen earns twice what Lorie earns, so in 3 hours she earns 2 x $30 = $60.
Karen works 3 hours per day, so in 2 days she works 3x 2 = 6 hours.
Therefore, in 2 days, Karen earns $60 x 6 = $360. - - → *It should be $60 x 2, rather than $60 x 6*
The answer is: 360

Illustration of diffusion intermediate process when the generative trajectory starts to be separated

token index 1 2 3 4 5 6 7 8 9 10 11
Baseline: Lorie earns $10 per hour, ⊗ ⊗ ⊗ ⊗ ⊗ ... - - - → *Step #7*

↓ *diffusion, do sampling procedure in Eq.(1)*

Top-1 confident index of single-step predictive distributions: 7th mask token

↓ *decoding*

Lorie earns $10 per hour, so ⊗ ⊗ ⊗ ⊗ ... - - - - → *Step #8*

Ours: Lorie earns $10 per hour, so Karen earns twice that, which is $10 x 2 = $20 per hour. - - - → *The generative trajectory starts to be different from the results of baseline at the "Karen"*
If Karen works 3 hours per day, she earns $20 x 3 = $60 per day.
In two days, Karen earns $60 x 2 = $120.
The answer is: 120

Illustration of diffusion intermediate process when the generative trajectory starts to be separated

token index 1 2 3 4 5 6 7 8 9 10 11
Ours: Lorie earns $10 per hour, ⊗ ⊗ ⊗ ⊗ ⊗ ... - - - - → *Step #7*

↓ *diffusion, do sampling procedure in Eq.(18)*

Top-1 confident index of approximated target distributions: 8th mask token

↓ *decoding*

Lorie earns $10 per hour, ⊗ Karen ⊗ ⊗ ⊗ ... - - → *Step #8, reject single-step top-1 confident index*

总结

1. **连贯上下文解码(CCD)：** 一种无需训练的推理框架，将DLM 采样重塑为一致性感知过程。
2. **轨迹矫正：** 利用边缘化上下文近似目标分布，提前拒绝次优的局部高置信度路径。
3. **自适应效率：** CCD-DS 打破了速度与精度的权衡，在Dream 和LLaDA 模型上实现了最高**3.48倍加速** 和**3.91% 的性能提升**。