用 SLOWFAST SAMPLING 加速扩散大型语言模型：三个黄金原则

Qingyan Wei[1,2,3] **Yaojie Zhang**[2,5] **Zhiyuan Liu**[1,2] **Dongrui Liu**[4] **Linfeng Zhang**[1,2*]
[1] School of Artificial Intelligence, Shanghai Jiao Tong University
[2] EPIC Lab, Shanghai Jiao Tong University
[3] Central South University
[4] Shanghai Artificial Intelligence Laboratory
[5] University of Electronic Science and Technology of China
{ florawei0506 } @gmail.com
Code: https://github.com/LiangrunFlora/Slow-Fast-Sampling

ABSTRACT

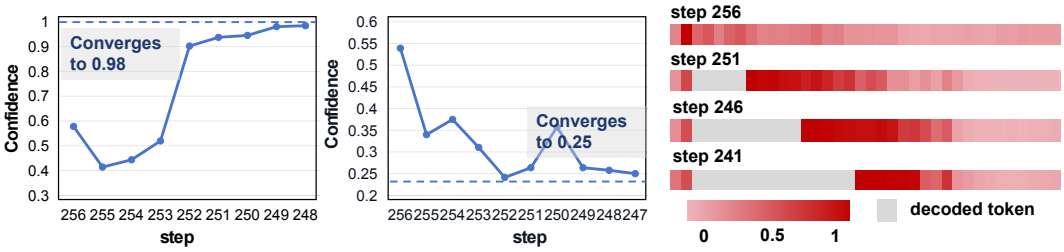
基于扩散的语言模型 (dLLMs) 通过支持并行的 token 生成和显著减少推理延迟，已经成为传统自回归大型语言模型的有前途的替代方案。然而，现有的 dLLMs 采样策略，例如基于置信度的或半自回归的解码，通常由于其静态行为而导致效率低下和灵活性有限。在本文中，我们提出了 SlowFast Sampling，这是一种新颖的动态采样策略，可以自适应地在探索性和加速解码阶段之间交替。我们的方法受到三条金原则的指导：确定性原则、收敛原则和位置原则，这些原则控制 token 何时何地可以被自信且高效地解码。我们进一步将我们的策略与 dLLM-Cache 集成，以减少冗余计算。在各种基准和模型上的广泛实验表明，SlowFast Sampling 在 LLaDA 上实现了最高 15.63 × 的加速，且准确性下降最小，当与缓存结合时，速度提升可达 34.22 × 。值得注意的是，我们的方法在吞吐量方面优于强大的自回归基线如 LLaMA3 8B，证明精心设计的采样策略可以释放 dLLMs 在快速和高质量生成方面的全部潜力。

1 介绍

大型语言模型（LLMs） (Zhao et al., 2025) 已迅速成为人工智能中的基石技术，展现出在不同自然语言理解和生成任务上的显著能力。然而，大多数 LLM 的普遍自回归特性，即令牌一个接一个地按顺序生成，使得推理延迟显著，特别是在处理长序列时。为了解决这个固有瓶颈，基于扩散的 LLM（dLLMs） (Ye et al., 2025; Nie et al., 2025b) 已成为一种有前景的替代范式。这些模型能够并行生成多个令牌，摆脱严格的逐令牌生成过程。这种并行解码能力提供了明显的优势，具有显著加速文本生成的潜力，使 dLLMs 成为高效语言模型推理的引人注目的前瞻性方向。然而，目前使用 dLLMs 进行采样的方法往往表现不如预期。常见的方法包括基于置信度的选择 (Chang et al., 2022)，在全球范围内选择置信度最高的固定数量令牌进行重新生成。另一种流行的方法，半自回归解码 (Arriola et al., 2025)，将序列划分为固定块并在其内进行解码。不幸的是，这些方法常常得到不理想的结果（即，当并行解码许多标记时，准确性显著下降），并且整个生成过程中具有静态、恒定的采样速度。缺乏灵活性突显了需要一种更加动态的采样方法：能够智能地决定在每一步采样多少标记，以及这些标记应该位于序列中的何处。受这些限制的启发，我们介绍了一种新颖的动态采样方法，旨在加速 dLLMs，旨在释放 dLLMs 在高并行解码下的真正潜力。如图 1 所示，我们的方法受到三个核心观察的指导，我们将其制定为有效加速的三个黄金原则，其如下：

- 确定性原则：表现出更高置信度的标记本质上更为确定。因此，它们更有可能在过程的早期被正确解码，并且在随后的扩散步骤中需要较少的调整。
- 收敛原理：随着扩散过程的展开以及符号的逐步优化，许多符号的语义意义趋于稳定，其相关的置信度分数也逐渐收敛到一个稳定值。这种收敛表明，这些符号已经基本定型，所需的进一步优化很少。

*Corresponding author.



(a) Confidence convergence diagram of the 1st token and the 12th token

(b) The Process of Change in the Confidence Map of LLaDA-Base

Figure 1: 扩散式 LLM 中采样的三个黄金原则的示意图。(a) 说明了收敛原则：随着解码的进行，标记的置信度值会收敛到稳定值，或高（例如，第 12 个标记达 0.98）或低（例如，第 1 个标记达 0.25）。(b) 可视化了在 LLaDA 基准的 GSM8K 上 256 个扩散步骤中的置信度图的演变。高置信度的标记（深红色） 逐步出现并优先解码（确定性原则），同时选择倾向于在连续区域中聚集（位置原则），从而实现缓存重用和高效加速。

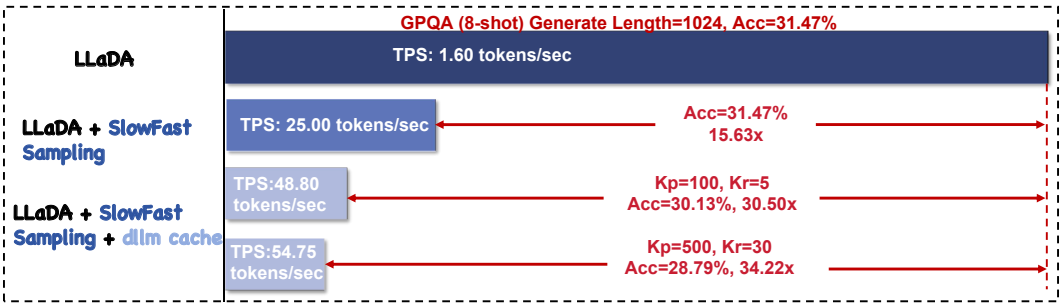


Figure 2: 在 GPQA（8-shot，长度 =1024）上使用 LLaDA 和我们提出的方法进行吞吐量和准确性比较。我们在三种设置下评估 LLaDA：（1）普通解码，（2）使用我们提出的慢快采样，以及（3）在慢快采样的基础上进一步增强的 dLLM-Cache。与普通设置相比，仅使用慢快采样即可实现 15.63 × 倍的加速，同时保持相当的准确性。使用 dLLM-Cache 后，吞吐量进一步提高到 54.75 tokens/sec（最高达到 34.22 × 倍的加速），且准确性仅有轻微下降。这表明我们的动态策略所实现的高效增益和灵活性。

- 位置原则：我们注意到，即使没有显性的约束，模型的采样偏好也往往倾向于特定位置的词元，这些位置通常是相邻或聚集的。这种内在的定位偏差可以被战略性地利用。例如，序列的部分可以被有效缓存，从而带来显著的加速效果。

结合这些原则，我们提出了慢速-快速采样方法，该方法分为两个不同的阶段：探索阶段和加速解码阶段。在最初的探索阶段，模型执行较不受约束的解码。这允许它自由探索序列空间。利用位置原则，模型识别出有前景的区域，并预测一个目标段，在此段中，标记显示出高不确定性和初步收敛。随后，加速解码阶段通过快速并行解码这些在识别出的段中高确定性、收敛的标记来利用这一点。通过集中计算努力在最有影响的地方，策略性地划分使得有效处理已大部分确定的标记，显著加速。如图 2 所示，我们的方法在 LLaDA 上实现了最大 15.63× 的加速。当进一步与 dLLM-Cache (Liu et al., 2025) 结合时，加速增加到令人印象深刻的 34.22× ，且精度仅有轻微下降。我们的贡献有三方面：

我们提出了基于标记确定性、收敛性和位置影响的三个黄金原则，这些原则在 dLLMs 的有效和高效采样中起关键作用。

在这些原则的基础上，我们引入了一种新的两阶段动态策略——SlowFast 采样，该策略专门设计用来利用这些原则以实现 dLLM 的最佳加速。

3. 通过在各种基准上的实验，我们展示了 SlowFast Sampling 在不影响响应质量的情况下实现了显著的推理加速（例如，单独使用 SlowFast Sampling 在 LLaDA 上可加速至 15.63 × ，与 dLLM-Cache 结合时可加速至 34.22 × ），因此与基线和更简单的采样方法相比，提供了更优的速度质量折中。

2 相关工作

2.1 语言的扩散模型

扩散模型（DMs）(Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) 在生成建模中引发了重大变革，尤其是在图像这样的连续领域中。但是，由于文本的离散性质，将这些模型适应于离散数据如文本时面临独特的挑战。在离散扩散模型中，一个有前景的方法涉及到掩码扩散模型（MDMs）(Austin et al., 2021; Lou et al., 2023; Shi et al., 2024; Nie et al., 2025a;b; Hoogeboom et al., 2021; Campbell et al., 2022)，它逐次根据上下文预测被掩码的标记。这些进展改变了文本生成，为大型语言模型（LLMs）中的自回归范式提供了一种引人注目的替代方案。值得注意的例子包括 LLaDA (Nie et al., 2025b)，一个从零开始用双向 Transformer 训练的 8B MDM，以及 Dream (Ye et al., 2025)，它从预训练的 ARM 权重初始化。这两个模型表现出与类似大小的 ARM 如 LLaMA3 8B (Dubey et al., 2024) 相当的性能。它们的双向架构可能克服 ARM 的限制，如反转诅咒 (Berglund et al., 2023)，使得扩散成为基础 LLMs 的一个有竞争力的替代方案。

由于 dLLMs 的高推理延迟，主要是因为其迭代去噪过程 (Nie et al., 2025b; Ye et al., 2025)，已促使对加速技术的研究。已经开发了各种策略，主要包括缓存机制和高级采样技术。

缓存机制。特征缓存通过在去噪步骤中重用中间特征来减少冗余计算。dLLM-Cache (Liu et al., 2025) 结合了长间隔提示缓存和短间隔响应缓存，利用 V-verify 机制有选择地更新动态令牌，从而显著提高推理速度。

高级采样技术。优化采样过程本身是加速 dLLMs 的另一个重要方向。低置信度重掩蔽 (Chang et al., 2022; Nie et al., 2025b) 优先考虑高置信度的标记以加速收敛；半自回归 (Nie et al., 2025b; Arriola et al., 2025) 重掩蔽将序列分成块，应用随机和低置信度策略。此外，MDMs 的精确模拟方法，如首次命中采样器 (Zheng et al., 2024)，在减少采样步骤或提高每步效率方面取得了进展。

然而，这些采样方法通常是静态的，缺乏灵活性。为了解决这一问题，我们提出了一种新的动态采样方法——SlowFast Sampling，通过整合确定性、收敛性和位置影响的原则，实现更高效的推理加速。

3 方法论

3.1 初步

扩散大语言模型的推理过程。扩散大语言模型（dLLMs）通过迭代去噪过程从提示 $\boldsymbol{c} = (c_1, \ldots, c_M)$ 生成文本 $\boldsymbol{y} = (y_1, \ldots, y_L)$。模型在离散步骤 N 上优化中间状态 $\boldsymbol{y}^{(k)} \in \mathcal{T}^L$，从 $k = N$ 到 $k = 0$，其中 $\mathcal{T}$ 是标记词汇表。该过程从一个完全掩码的序列开始：

$$\boldsymbol{y}^{(N)} = (\underbrace{\texttt{[MASK]}, \ldots, \texttt{[MASK]}}_{L \text{ times}}) \tag{1}$$

其中 $\texttt{[MASK]} \in \mathcal{T}$ 是特殊掩码标记。

在每个步骤 $k \in \{N, N-1, \ldots, 1\}$ 中，掩码预测器 p_θ 从噪声状态 $\boldsymbol{y}^{(k)}$ 和提示 $\boldsymbol{c}$ 估计原始干净序列 $\boldsymbol{r}_0 = (r_{0,1}, \ldots, r_{0,L})$：

$$P_\theta(\boldsymbol{r}_0 | \boldsymbol{c}, \boldsymbol{y}^{(k)}) \tag{2}$$

中的干净序列的估计值通过贪心解码在步骤 k 中获得，$\hat{\boldsymbol{r}}_0^{(k)}$：虽然预测器 p_θ 可以在一个步骤中解码所有被掩盖的标记 $\texttt{[MASK]}$，但为了确保高质量的生成，dLLM 采用多步骤解码过程。在每个步骤中，重新掩码策略优化标记。过渡到下一个状态 $\boldsymbol{y}^{(k-1)}$ 由采样策略 S 控制：这个迭代去噪过程是一个采样程序。我们的工作旨在提高 dLLM 推理的采样效率，由于 N 顺序步骤可能会导致计算量大。以 LLaDA (Nie et al., 2025b) 为例，我们描述其核心采样策略。在 LLaDA 中，时间步骤被定义为 $t_k = k/N$ 和 $t_{k-1} = (k-1)/N$。

LLaDA 探索了在方程式 **??** 中转换函数 S 的各种策略，主要区别在于来自 $\hat{\boldsymbol{r}}_0^{(k)}$ 的标记如何更新 $\boldsymbol{y}^{(k)}$ 以形成 $\boldsymbol{y}^{(k-1)}$，特别是对于在 $\boldsymbol{y}^{(k)}$ 中为 $\texttt{[MASK]}$ 的位置：

随机重新遮盖。在此策略中，对于每个位置 i：
如果 $y_i^{(k)} \neq \texttt{[MASK]}$，则 $y_i^{(k-1)} = y_i^{(k)}$（已知的令牌被保留）。

如果 $y_i^{(k)} = \texttt{[MASK]}$，则以概率 $1 - \frac{k-1}{k}$ 将 $y_i^{(k-1)}$ 设置为 $\hat{r}_{0,i}^{(k)}$，并以概率 $\frac{k-1}{k}$ 保持 [MASK]，确保遮盖的令牌的期望数量与噪声计划一致。

低置信度重掩盖。这种确定性策略旨在通过选择性地解除标记掩盖来提高样本质量。对于每个位置 i，如果 $y_i^{(k)} = \texttt{[MASK]}$，模型会预测 $\hat{r}_{0,i}^{(k)}$ 并计算其置信度。具体来说，置信度 c_i 由以下式给出：

$$c_i = P_\theta(\hat{r}_{0,i}^{(k)} | \boldsymbol{c}, \boldsymbol{y}^{(k)}). \tag{3}$$

如果 $y_i^{(k)} \neq \texttt{[MASK]}$，则 $c_i = 1$。

状态 $\boldsymbol{y}^{(k-1)}$ 的目标未掩盖标记数由以下公式给出：

$$n_{un} = \lfloor L(1 - t_{k-1}) \rfloor = \lfloor L\left(1 - \frac{k-1}{N}\right) \rfloor. \tag{4}$$

。对应于最高置信度 n_{un} 的标记在 $\boldsymbol{y}^{(k-1)}$ 中未被掩盖，而其余位置则设为 [MASK]。

半自回归再掩盖。这种策略在监督微调后用于 LLaDA。序列被划分为若干块，然后从左到右逐块生成。在每个块中，在移到下一个块之前，迭代地应用随机再掩盖或低置信度再掩盖策略来对块进行去噪。

采样策略 S 的选择以及步骤数量 N 显著影响生成质量和延迟。

3.2 dLLM 的三个黄金原则

在第 3.1 节中概述的迭代去噪过程基础上，我们对 dLLM 行为的实证分析，特别是像 LLaDA 这样的模型，揭示了在生成标记时的一致模式。我们称这些模式为三个黄金原则，它们构成了我们提出的加速策略的基石。它们描述了在扩散过程中标记的确定性、收敛性和位置效应是如何相互作用的，从而为优化采样提供了关键见解。

确定性原则：高置信度表明决心。我们观察到，蒙版预测器 p_θ 以高置信度预测的标记显著更有可能成为最终正确序列的一部分，并且在后续去噪步骤中趋于保持不变。在位置 i 和步骤 k 预测的标记 $\hat{r}_{0,i}^{(k)}$ 的置信度由 $P_\theta(\hat{r}_{0,i}^{(k)} | \boldsymbol{c}, \boldsymbol{y}^{(k)})$ 给出。如图 1 (b) 所示，在任意给定的步骤 k 中，通常有一部分标记表现出显著高于其他标记的置信得分。这些高置信度的标记是" 提前接受" 或较少重新采样的理想候选者。

加速的意义：通过优先处理快速达到高置信度阈值的标记，我们可以减少对序列中已确定部分的冗余计算。

收敛原则：迭代过程中令牌稳定。在迭代细化过程中，个别令牌（无论是预测的身份 $\hat{r}_{0,i}^{(k)}$ 还是它们的置信度 $P_\theta(\hat{r}_{0,i}^{(k)} | \boldsymbol{c}, \boldsymbol{y}^{(k)})$）都会经历一段波动期，然后才稳定。如图 1 所示，一个代表性的令牌可能在几个早期扩散步骤中最初改变其预测的身份和置信度。然而，随着 k 的减少，令牌的信心常常收敛到一个稳定值。这种收敛表示模型在当前上下文中形成了对令牌身份的一致信念。

加速的意义：已经显示出收敛的标记（即，在最近几步的窗口中呈现稳定的身份和置信度）不太可能发生变化。积极的解码可以防止不必要的重新评估，从而加快过程。

位置原则: 解码表现区域偏好。除了个别标记行为之外，我们发现生成过程通常表现出空间模式。高置信度和早期收敛的标记不会随机分布在整个序列中。相反，它们经常以连续的块或局部区域的形式出现，如图 1 (b) 所示。这种现象可能是由于局部语义依赖性或提示 $\boldsymbol{c}$ 的强烈上下文化部分的影响。例如，在一些初始步骤之后，某个特定的标记跨度可能集体达到高置信度和稳定性，而其他区域则仍然大部分被掩盖或不确定。模型似乎在不同阶段将其解码努力集中在特定的片段上。

对加速的启示：识别这些区域可以实现有针对性的解码。与其对所有标记进行统一处理，可以将计算资源集中在当前最适合解码的区域。

这三条原则共同表明，一刀切的静态采样策略本质上是低效的。它们激励了一种动态方法，其中采样的 token 数量、选择标准和位置根据生成序列的变化状态进行调整。我们的 SlowFast 采样方法，详见第 3.3 节，旨在明确利用这些观察结果以实现显著的推理速度提升。

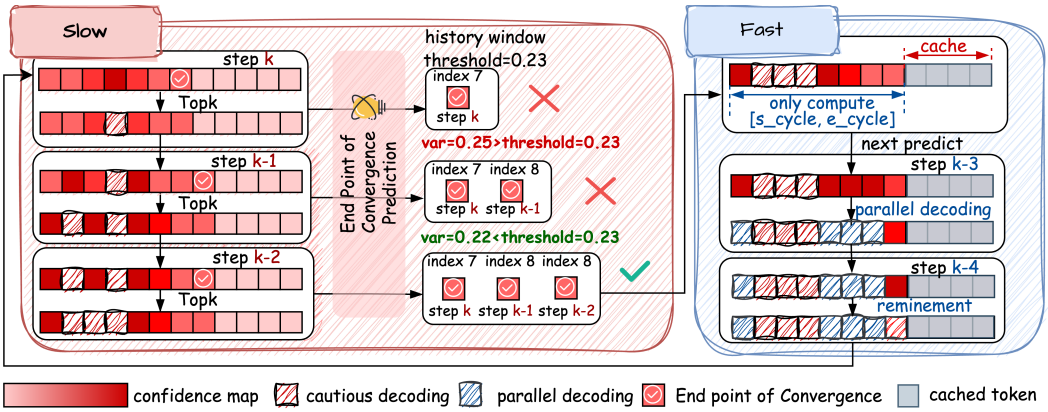


Figure 3: 慢-快采样流水线概述：从探索到加速解码。该方法在慢（探索）阶段和快（加速）阶段之间交替，以实现高效的标记生成。在慢阶段（左），模型通过在每一步选择前 k 个高置信度的标记进行谨慎解码，同时持续预测收敛终点并计算历史窗口内的置信度方差。一旦方差降至阈值（例如，$0.22 < 0.23$）以下，相应区域 $[s_{cycle}, e_{cycle}]$ 被视为稳定。在快阶段（右），此稳定区间通过积极地去除高置信度标记的掩蔽进行并行解码，而区间之外的标记则暂时跳过，其结果被缓存以供重用。这种交替结构减少了冗余计算，加速了解码，同时保持输出质量。

3.3 慢-快采样

1. 探索阶段（慢速阶段）：识别下一个稳定区域。如图 3 所示，此阶段的主要目标是在解码时谨慎前进，同时识别一个有希望的、稳定的区域，以进行后续的快速处理。从 s_{cycle} 开始并延伸至完整序列的结束 L，此阶段在有限数量的 dLLM 步骤中如此操作：

一旦满足这一稳定性标准，探索阶段即告结束。后续加速解码阶段的终点 e_{cycle} 被设定为最后记录的候选地平线 $e^{(k)}_{cand}$。如果在最大探索步骤数内未达到稳定性，e_{cycle} 可以设为保守确定的位置，或者该过程可能会默认进行该周期的完整序列谨慎解码。

2. 加速解码阶段 (快速阶段): 快速并行优化。如图 3 的右半部分所示，一旦识别出稳定区域 $[s_{cycle}, e_{cycle}]$，此阶段旨在快速去噪该范围内的标记，同时有效处理该范围之外的标记：

- 超出范围缓存：对于超出识别范围的标记（即，位置 $i > e_{cycle}$），如果它们当前预测的置信度较低（例如，低于 τ_{min_conf}），则在该阶段的一个 dLLM 步骤中计算它们的预测值 $\hat{r}^{(k)}_{0,i}$，并进行缓存。这些缓存值可以在后续的 dLLM 步骤中复用于这些位置，前提是它们仍然位于活动解码范围之外，从而节省冗余计算。
- 区间内并行解码：在区间 $[s_{cycle}, e_{cycle}]$ 内，尝试进行激进的并行解码。所有对于 $i \in [s_{cycle}, e_{cycle}]$ 的 $y^{(k)}_i = \texttt{[MASK]}$ 令其预测置信度 $P_\theta(\hat{r}^{(k)}_{0,i} | \boldsymbol{c}, \boldsymbol{y}^{(k)})$ 超过高置信度阈值 τ_{high_conf} 的，即

$$P_\theta(\hat{r}^{(k)}_{0,i} | \boldsymbol{c}, \boldsymbol{y}^{(k)}) > \tau_{high\_conf} \tag{5}$$

  的值被设置为对应的预测 $\hat{r}^{(k)}_{0,i}$ 来取消掩码。此更新对于所有符合条件的 token 在一个概念步骤中同时执行，以形成 $\hat{\boldsymbol{r}}^{(k)}_0$。
- 回退 Top-k 优化：如果在给定范围内满足 τ_{high_conf} 标准的标记数量不足以取得显著进展（例如，少于一个），我们将在该范围内的这一 dLLM 步骤中退回到一个更保守的更新。具体来说，我们根据置信度选择 $[s_{cycle}, e_{cycle}]$ 中排名前 k_{fast} 的标记，并将其解码。即使未达到广泛的高置信度，这仍能确保稳步的进展。

加速解码阶段完成后，下一周期的探索阶段的起始位置将更新为 $s_{cycle} \leftarrow e_{cycle}$。这种探索和加速解码的循环过程持续进行，直到整个序列生成完毕。此“慢快”方法动态调整解码的焦点和强度，利用确定性和位置原则识别有潜力的区域，并通过收敛原则高效稳定这些区域。

Table 1: LLaDA 8B 和 Dream 7B 在 8 个基准测试中与 SlowFast Sampling 的性能表现。

Task	Method	Inference Efficiency		Performance	Method	Inference Efficiency		Performance
		TPS ↑	Speed(TPS) ↑	Score ↑		TPS ↑	Speed(TPS) ↑	Score ↑
Mathematics & Science								
GSM8K	LLaDA 基础	4.55	1.00 ×	69.83	梦想基地	8.16	1.00 ×	77.02
	+ Sampling	14.57 +10.02	3.20 ×+2.20	69.59 −0.27	+ Sampling	17.15 +8.99	2.10 ×+1.10	76.50 −0.52
GPQA	LLaDA 基础	3.31	1.00 ×	31.47	梦境基地	5.43	1.00 ×	35.93
	+ Sampling	16.36 +13.05	4.94 ×+3.94	31.91 +0.44	+ Sampling	16.56 +11.13	3.05 ×+2.05	35.94 +0.01
Math	LLaDA 基础	5.14	1.00 ×	30.16	梦想基地	8.48	1.00 ×	38.68
	+ Sampling	11.27 +6.13	2.19 ×+1.19	29.64 −0.52	+ Sampling	23.00 +14.52	2.71 ×+1.71	38.24 −0.44
General Tasks								
MMLU-pro	LLaDA Base	9.16	1.00 ×	23.30	梦想基地	14.97	1.00 ×	24.14
	+ Sampling	23.14 +13.98	2.53 ×+1.53	23.85 +0.55	+ Sampling	22.80 +7.83	1.52 ×+0.52	22.91 −1.23
MMLU	LLaDA 基础	5.02	1.00 ×	62.11	梦想基地	8.46	1.00 ×	72.61
	+ Sampling	16.81 +11.79	3.35 ×+2.35	66.56 +4.45	+ Sampling	18.43 +9.97	2.18 ×+1.18	75.13 +2.52
BBH	LLaDA 基础	4.04	1.00 ×	44.97	梦基地	6.93	1.00 ×	51.83
	+ Sampling	21.19 +17.15	5.24 ×+4.24	44.60 −0.37	+ Sampling	28.14 +21.21	4.06 ×+3.06	50.55 −1.28
Code								
MBPP	LLaDA Base	4.98	1.00 ×	40.80	梦想基地	8.92	1.00 ×	54.20
	+ Sampling	13.32 +8.34	2.67 ×+1.67	41.00 +0.20	+ Sampling	29.07 +20.15	3.26 ×+2.26	54.60 +0.40
HumanEval	LLaDA 基础	11.24	1.00 ×	31.71	梦境基础	11.49	1.00 ×	54.26
	+ Sampling	35.46 +24.22	3.15 ×+2.15	33.54 +1.83	+ Sampling	25.38 +13.89	2.21 ×+1.21	52.43 −1.83

Table 2: LLaDA Base 与 SlowFast Sampling 和 dLLM-Cache 的性能表现。

Task	Method	Inference Efficiency		Performance
		TPS ↑	Speed(TPS) ↑	Score ↑
Mathematics & Science				
GSM8K	LLaDA 基础	4.55	1.00 ×	69.83
	Sampling + Cache	26.99 +22.44	5.83 ×+4.83	69.60 −0.23
GPQA	LLaDA 基础	3.31	1.00 ×	31.47
	Sampling + Cache	29.06 +25.75	8.78 ×+7.78	33.48 +2.01
Math	LLaDA 基础	5.14	1.00 ×	30.16
	Sampling + Cache	26.50 +21.36	5.16 ×+4.16	29.42 −0.74
General Tasks				
MMLU-pro	LLaDA 基础	9.16	1.00 ×	23.30
	Sampling + Cache	33.38 +24.22	3.64 ×+2.64	25.53 +2.23
MMLU	LLaDA Base	5.02	1.00 ×	62.11
	Sampling + Cache	38.42 +33.40	7.65 ×+6.65	61.20 −0.91
BBH	LLaDA 基础模型	4.04	1.00 ×	44.97
	Sampling + Cache	36.04	8.92 ×+7.92	44.81 −0.16
Code				
MBPP	LLaDA Base	4.98	1.00 ×	40.80
	Sampling + Cache	27.26 +22.28	5.47 ×+3.87	39.00 −1.80
HumanEval	LLaDA Base	11.24	1.00 ×	31.71
	Sampling + Cache	41.14 +29.90	3.66 ×+2.66	31.10 −0.61

4 实验

4.1 实验设置

实施细节为了评估我们提出的动态采样方法 SlowFast Sampling 的有效性，我们在具有代表性的 dLLMs 上进行了实验：LLaDA 8B (Nie et al., 2025b) 和 Dream 7B (Ye et al., 2025)，重点是测量在各种基准测试中的推理加速。所有实验均在 NVIDIA RTX 4090 GPUs 上进行。

评估指标我们使用定量指标评估 SlowFast Sampling 的采样加速和生成质量。推理速度以每秒生成的令牌数（TPS）来衡量，表示每秒生成的平均令牌数量。生成质量使用特定任务指标进行评估，例如 GSM8K 的准确性，反映了模型在推理加速下的性能。

Sampling Strategy	TPS ↑	Score ↑
Autoregressive (AR)	5.25	60.80
Diffusion Sampling	4.55	69.83
Semi-Autoregressive	5.44	66.41
SlowFast Sampling	9.87	69.59

Table 3: Comparison of Sampling Strategies on inference efficiency and generation performance.

Method	TPS ↑	Speed(TPS) ↑	Accuracy ↑
LLaMA3 8B (Dubey et al., 2024)	33.79	1.00 ×	31.92
LLaDA Base	1.60 −32.19	1.00 ×	31.47 −0.45
+ Sampling	25.00 −8.79	15.63 ×	31.47 +0.00
+ Sampling + Cache ($K_p = 100, K_r = 5$)	48.80 +15.01	30.50 ×	30.13 −1.34
+ Sampling + Cache ($K_p = 500, K_r = 30$)	54.75 +20.96	34.22 ×	28.79 −3.13

Table 4: Comparison of LLaDA 8B Base with other representative LLMs . Compared to LLaMA3 8B, LLaDA with SlowFast Sampling and dLLM-Cache achieves significantly higher throughput (up to +20.96 TPS) while maintaining comparable accuracy.

4.2 主要结果

各模型的性能和效率提升表 1 报告了 LLaDA 8B 和 Dream 7B 的吞吐量和模型性能，分别在有和没有 SlowFast Sampling 的情况下。这些结果表明，我们的方法在推理效率上显著提升，并且在大多数情况下实现无损加速。在我们的实验中，SlowFast Sampling 、τ_{min_conf} 和 τ_{high_conf} 的关键超参数分别设置为 0.1 和 0.85 。我们方法中的其余超参数设置如下：最大探索步数 $K_{max} = 8$ 、滑动窗口大小 $W_{hist} = 2$ 、和稳定方差阈值 $\sigma^2_{stable} = 1.0$ ，与第 4.2 节和消融研究中使用的默认配置一致。

与 dLLM-Cache 的兼容性。最近，几项研究探讨了利用特征缓存来降低 dLLM 推理的计算成本。我们的 SlowFast Sampling 与现有的缓存机制高度兼容，这种集成与单独使用相比可以带来更高的加速。表 2 比较了在结合使用我们的方法和 dLLM-Cache 与未应用此方法的情况下，LLaDA 的性能和推理速度。结果表明，此集成可以在保持可比模型性能的同时提供更高的吞吐量。

与其它采样策略的比较我们将我们的 SlowFast Sampling 方法与三种替代采样策略进行了比较：扩散采样、扩散半自回归采样和自回归（AR）采样。扩散采样采用重新掩码策略以迭代方式并行选择要解码的标记，而 AR 采样严格从左到右生成标记。半自回归方法从左到右生成块，并在每个块内应用重新掩码策略。如表 3 所示，我们的方法 SlowFast Sampling 在保持竞争性生成质量的同时，达到更高的推理效率。

4.3 消融研究

最小置信度阈值 τ_{min_conf} 和高确定性阈值 τ_{high_conf} 的效果。

核心置信度阈值 τ_{min_conf} 和 τ_{high_conf} 决定了我们 SlowFast 管道的行为。如图 4 所示，最小置信度阈值 τ_{min_conf} 定义了探索阶段的候选区域。$\tau_{min\_conf} = 0.1$ 的中等值是最佳的，因为较低的值会导致不稳定的区域，而较高的值会导致过于保守、效率低下的区域。在加速阶段，高确定性阈值 τ_{high_conf} 直接管理速度与质量的权衡。较高的值会导致更谨慎和准确的生成，但以速度（TPS）为代价。我们选择了 $\tau_{high\_conf} = 0.85$ ，它在不显著降低推理速度的情况下实现了接近峰值的 GSM8K 分数，从而实现了有效的平衡。

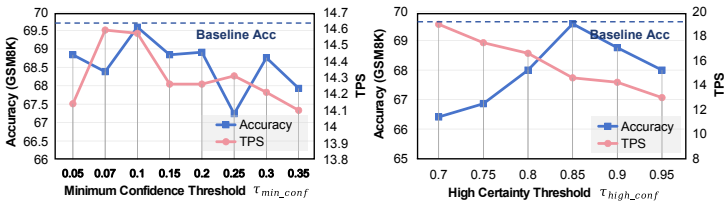


Figure 4: 置信阈值的影响。τ_{min_conf} 控制探索范围，τ_{high_conf} 在快速解码过程中平衡准确性和速度。其他超参数遵循在第 4.2 节中的默认设置。

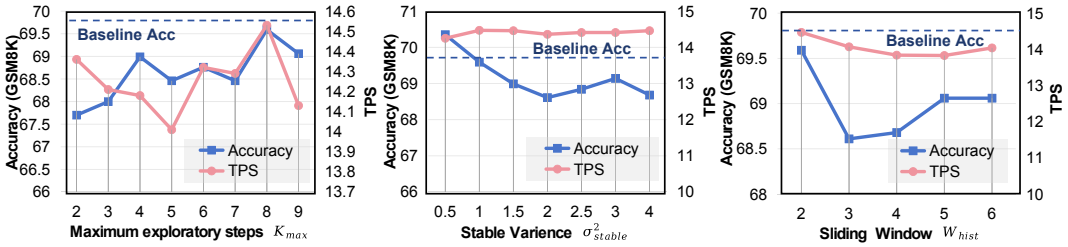


Figure 5: 稳定性检查中超参数的敏感性研究。准确性和 TPS 会随 K_{max} 、σ^2_{stable} 和 W_{hist} 而变化。选择的默认值（$K_{max} = 8$ 、$\sigma^2_{stable} = 1.0$ 、$W_{hist} = 2$ ）提供了强有力的速度和质量权衡。

超参数在稳定性检测中的作用。通过三个超参数的调节，稳定性检测使模型从慢速阶段过渡到快速阶段，其效果如图 5 所示。探索步数的最大值 K_{max} 对于使收敛范围稳定至关重要；我们发现 $K_{max} = 8$ 提供了足够的探索以实现高质量生成而不会产生过多的开销。滑动窗口大小 W_{hist} 也对此起到补充作用。由于预测变化和收敛发生得很快，我们发现使用较小的窗口 $W_{hist} = 2$ 能够有效地在保持质量和最大化推理速度之间取得良好平衡。最后，严格的稳定方差阈值 $\sigma^2_{stable} = 1.0$ 确保加速阶段仅在真正稳定的区域才会触发，以此巩固我们方法的可靠性。

在推理速度上超越自回归 LLMs。如表 2 所示，当配备我们的 SlowFast 采样和 dLLM 缓存时，LLaDA Base 不仅在其默认设置上显著加速，还在推理速度上超越了自回归 LLaMA3 8B 模型 Dubey et al. (2024) （54.75 对 33.79 TPS），同时保持了相当的准确性。这表明 dLLMs，通过适当的采样和优化，可以在效率和实用性上超越传统的自回归模型。

5 结论

在这项工作中，我们提出了 SlowFast 采样，一种动态且有原则的方法，用于加速基于扩散的大型语言模型（dLLMs）。通过利用三个关键观察：符号确定性、收敛性和位置偏差，我们设计了一个两阶段解码管道，自适应地平衡探索和高效的并行解码。在基准测试中的大量实验表明，我们的方法不仅显著提高了推理速度（最高可达 **15.63** × 和与 dLLM-Cache 结合使用时最高可达 34.22 × ），而且也保持了强大的生成质量，甚至在吞吐量上超过了自回归 LLMs。我们相信这项工作标志着使 dLLMs 在实际部署中切实可行且具有竞争力的重要一步。

REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Lukas Berglund, Meg Tong, Max 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". *arXiv preprint arXiv:2309.12288*, 2023.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyan Wei, Shaobo Wang, and Linfeng Zhang. dllm-cache: Accelerating diffusion large language models with adaptive caching, 2025. URL https://github.com/maomaocun/dLLM-cache.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion language modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text, 2025a. URL https://arxiv.org/abs/2410.18514.

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, 2025b. URL https://arxiv.org/abs/2502.09992.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025. URL https://hkunlp.github.io/blog/2025/dream.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. URL https://arxiv.org/abs/2303.18223.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.