为最坏做准备，为最好做规划：理解掩码扩散中的标记排序

金佳妍[*1] 库林·沙阿[*2] 瓦西里斯·孔托尼斯[2] 沙姆·卡卡德[1] 陈思坦[1]

摘要

近年来，掩码扩散模型（MDMs）作为生成建模在离散领域中的新方法崭露头角。相较于自回归模型（ARMs），MDMs在训练时需要处理指数级规模的填充问题，但在推理时却能以灵活的方式解码标记。本研究深入探讨了这两种相互制约的特性：在训练阶段，我们通过理论与实证分析证明，MDMs确实需要处理比自回归模型更难的计算子问题；而在推理阶段，通过自适应选择标记解码顺序的策略，MDMs能有效规避复杂子问题，显著提升模型性能。在数独等逻辑谜题中，我们证明自适应推理能将预训练MDMs的解题准确率从不足7%提升至≈ 90%，甚至超越参数量多出7 ×且通过教师强制训练明确学习解码正确顺序的ARMs模型。这表明在训练和推理过程中未掌握正确标记生成顺序的MDMs，其表现仍可超越基于正确顺序知识训练的ARMs。我们还验证了自适应 MDM 推理在80亿语言扩散模型（LLaDa 8B）上的编码、数学等推理任务中的有效性。

[*]同等贡献[1]哈佛大学[2]德克萨斯大学奥斯汀分校。通讯作者：Kulin Shah <kulin shah@utexas.edu>。

第42届国际机器学习会议论文集，加拿大温哥华。 PMLR 267,2025。

1. 介绍

虽然扩散模型（Ho等人，2020；Song等人，2021）如今已成为图像、视频和音频等连续域生成建模的主流方法，而将该方法扩展到文本和蛋白质等离散域的研究（Austin 等人，2021；Lou 等人，2024；Hoogeboom等人，2021b）仍处于萌芽阶段。在众多提案中，掩码扩散模型（MDMs）（Lou 等人，2024；Sahoo等人，2025；Shi 等人，2024）已崭露头角，成为领先方案，其独特之处在于一个简单而原则性的目标：生成样本，学习逆转独立随机掩码标记的噪声过程。

在许多应用中，例如语言建模，掩码扩散模型（MDMs）与自回归模型（ARMs）相比仍表现欠佳（Nie 等人，2024；Zheng等人，2024），后者通过学习逆向噪声过程，从左到右依次解掩码标记。然而，近期研究表明，MDMs可能在ARMs表现不足的领域具有优势，包括推理（Nie 等人，2024；Kitouni 等人，2025）、规划（Ye 等人，2024）以及填充（Gong 等人，2024）。这引发了一个关键问题：与ARMs相比，MDMs的优势和局限性是什么？在何种类型的任务中，MDMs能够扩展以挑战ARMs在离散生成建模中的主导地位？

为理解这些问题，我们在权衡MDMs与ARMs的优劣时，将显微镜对准两个关键竞争因素：

- **训练时的复杂性**：MDM（多维向量模型）在设计上面临更具挑战性的训练任务。与仅需预测未掩码前缀的下一个标记的ARM（自编码器）不同，MDM需要根据任意位置的未掩码标记集合来预测标记，这本质上增加了其训练复杂度。

- **推理时的灵活性**：另一方面， MDM 采用的采样路径更为灵活。与ARMs固定的从左到右解码方式不同，MDMs在推理时会随机解码令牌。更有可能的是：MDM可以*以任何顺序*（包括从左到右）进行解码。

因此，我们提出以下问题：

MDM的推理灵活性优势是否足以抵消训练复杂性的弊端？

在本研究中，我们针对该问题提供了双重视角。

(1) 为最坏情况做准备。首先，我们通过理论与实证研究证明：训练复杂度带来的额外开销会量化影响MDM模型的性能表现。从理论上讲，我们展示了具有自然左向右顺序的简单数据分布案例，其中ARM模型能够高效生成样本。相比之下，在某些噪声水平下，MDM模型针对这些分布求解的子问题中，有相当大比例被证明存在计算不可行性。通过实证分析，我们在具有左向右顺序特征的真实文本数据上验证了这一结论，并发现即使在真实文本数据中，不同子问题间的训练复杂度失衡现象依然存在（图2，左）。

(2) 为最佳状态做规划。虽然上述内容看似对MDM不利，但在本文第二部分中，我们通过基于以下观察（Chang 等人，2022；Zheng 等人，2023）得出肯定回答：能够完美解决所有掩码子问题的MDM可用于*任意*顺序的解码。

在论文的第一部分，我们发现MDM模型训练过程中子问题间的复杂度失衡会导致部分子问题训练不足，而采用常规 MDM 推理（即随机顺序解码标记）会评估这些训练不足的边缘分布。因此，我们采用*自适应*策略替代常规 MDM 推理，通过智能选择解码顺序来优化训练效果。我们的核心发现是：这种自适应策略能够有效*离开*难度较高的子问题进行训练（见图1）。特别值得注意的是，即使**不改变MDM模型的训练方式**，最终模型的logit输出仍包含足够的信息来确定正确的解码顺序。我们展示了自适应推理在解决逻辑谜题、编码、数学和填空任务中的有效性。例如，在数独谜题中，一种简单的自适应策略（第4.1节）将MDMs的准确率从<7%提升至近90%。

MDMs相较于ARMs的优势。我们发现MDMs的主要优势体现在需要跨序列处理非*固定*自然词序的任务场景（例如逻辑谜题、编程和数学推理等）。通过精心设计逻辑谜题实验，我们证明了**在训练和推理阶段未掌握正确词序知识的MDMs**，其表现优于**掌握正确词序知识训练的ARMs**。特别值得注意的是，我们发现通过自适应策略在推理过程中决定正确标记生成顺序的MDMs，可以优于通过监督式教师强制训练来学习正确标记生成顺序的ARMs（Shah 等人，2024；Lehnert 等人，2024）。

组织。在第2节中，我们介绍了MDMs的基础知识和集合表示法。第3节探讨了 MDM 训练，并展示了不同子问题间计算不可行性的不平衡性。第4节研究了MDMs中的自适应推理，并分析了其对各类任务似然建模的影响。

2. 掩码扩散模型（Masked Diffusion Models，简称 MDM）

在本节中，我们解释了Masked Diffusion Models（Shi 等人，2024；Sahoo 等人，2025）的框架，并强调其作为*无序学习器*的解释。MDMs逐步向真实离散数据添加掩码噪声，并学习诱导逆过程的边缘分布。我们在下文正式定义MDMs的前向和逆向过程。

设分布p_{data}在$\{1, \ldots, m\}^L$上是长度为L且词汇表为$\{1, \ldots, m\}$的序列数据分布。我们用0表示“掩码”标记。

正向过程。对于给定的$x_0 \sim p_{\text{data}}$和噪声水平$t \in [0, 1]$，前向过程$x_t \sim q_{t|0}(\cdot|x0)$是通过$q_{t|0}(x_t|x_0) = \prod_{i=0}^{L-1} q_{t|0}(x_t^i|x_0^i)$, where

$$q_{t|0}(xit|xi0) = \text{Cat}(\alpha_t \mathbf{e}_{x_{i0}} + (1 - \alpha_t)\mathbf{e0})。$$

此处，α_t是一个预定义的噪声调度，满足$\alpha_0 \approx 1$、$\alpha_1 \approx 0$，且$\mathbf{e}_{x_{i0}} \in \mathrm{R}^{m+1}$是对应于标记$xi0$值的独热向量。Cat（$\pi$）表示由$\pi \in \Delta^m$给出的分类分布。换言之，对于每个第$i$个坐标，$xit$以概率 $1 - \alpha_t$被掩码为掩码标记 0，其余部分保持不变。

反向过程。上述正向过程的逆过程记为$q_{s|t}(x_s|x_t, x0)$，其表达式为$q_{s|t}(x_s|x_t, x_0) = \prod_{i=0}^{L-1} q_{s|t}(xis|x_t, x0)$对于任意$s < t$，其中

$$q_{s|t}(x_s^i \mid x_t, x_0) = \begin{cases} \text{Cat}(\mathbf{e}_{x_t^i}) & x_t^i \neq 0 \\ \text{Cat}\left(\frac{1-\alpha_s}{1-\alpha_t}\mathbf{e}_0 + \frac{\alpha_s-\alpha_t}{1-\alpha_t}\mathbf{e}_{x_0^i}\right) & x_t^i = 0\,. \end{cases}$$

反向转移概率$q_{s|t}(x_s^i|x_t, x0)$可通过$g_\theta(xis|xt)$。$q_{s|t}(xis|x_t, x_0 \leftarrow p_\theta(\cdot|x_t, t))$进行近似，其中$p_\theta(\cdot|x_t, t)$是一个去噪网络，通过基于 ELBO 的损失函数训练以预测噪声尺度t下所有掩码标记在$xi0$上的边缘分布（即对于所有满足条件的i

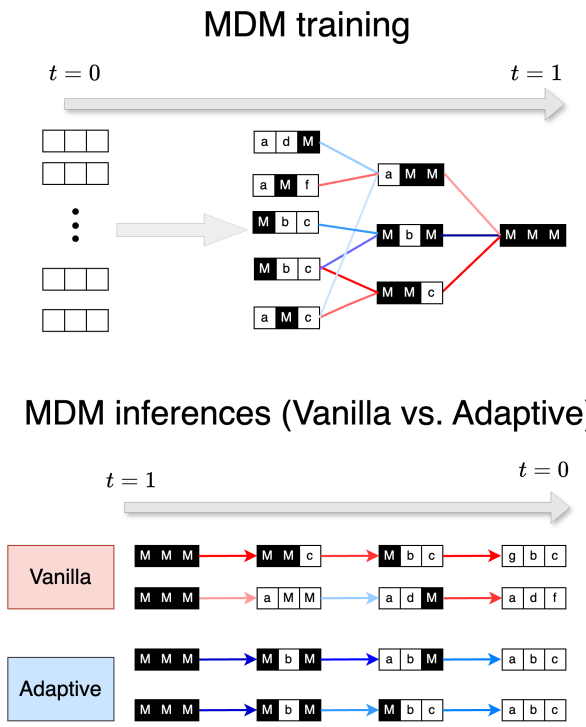


图1.（上）MDM 训练可视为学习多个掩码预测问题，其中部分问题更难学习，导致性能失衡（第3节）。（下）在推理过程中，自适应 MDM 可避免困难问题实例，从而提升性能（第4节）。

$xit=0$）。具体而言，$q_{s|t}$（$xis\,|x_t, x_0 \leftarrow p_\theta(\cdot|x_t, t)$）表示条件概率，其中 $p_\theta(\cdot|x_t, t)$ 被置于 $q_{s|t}$（$xis\,|x_t, x0$）内of $e_{x\,i0}$位置。去噪网络通过最小化以下损失函数进行训练，该损失源自分数熵（Lou 等人，2024；Sahoo 等人，2025；Shi等人，2024；Ou等人，2024）：

$$\mathcal{L}_\theta = \int_0^1 \frac{\alpha_t'}{1-\alpha_t} \mathop{\mathbb{E}}_{\substack{x_0\sim p_{\text{data}} \\ x_t\sim q_{t|0}(\cdot|x_0)}} \sum_{i:x_t^i=0} -\log p_\theta(x_0^i|x_t,t)dt,$$

在哪里$\alpha_t' = \frac{d\alpha_t}{dt}$ 求和运算是在掩码标记上进行的（即所有i，其中$x_t^i = 0$). 在实践中，通常采用一种无需时间嵌入的去噪网络架构，即$p_\theta(\cdot|x_t, t) = p_\theta(\cdot|xt)$，因为$x_t$通过掩码标记的数量隐式包含了关于$t$的信息。

反向采样过程从完全掩码的句子x_1=（0，...，0）开始。假设我们在给定噪声水平$t \in (0,1]$下有一个部分x_t被完全掩码的序列。那么，为了在预定噪声水平$s < t$下获得x_s，我们需要对所有i进行$xis \sim g_\theta(\cdot|xt)$的采样。该过程从$t$= 1递归重复到$t$= 0。

2.1. MDM的训练与推理重构

在本节中，我们首先讨论MDM的训练，并将其与2.1.1节中自回归模型的“从左到右”顺序训练进行比较。然后，我们在2.1.2节中重新定义了vanilla MDM 推理，为接下来的讨论奠定基础。

2.1.1. MDMS 的无序训练

近期研究（Zheng 等人，2024；Ou 等人，2024）发现 MDM 的学习问题等价于掩码语言模型。基于他们的分析，我们重新定义损失L_θ，证明L_θ是所有可能填充掩码损失的线性组合。我们首先将$x_0[M]$定义为掩码序列，该序列由原始序列x_0通过将掩码集M（$[L]$。{1, 2，...，L}的子集）中的索引替换为掩码标记0得到。

命题2.1。 *假设* α_0= 1，α_1= 0*且去噪网络* p_θ *不含时间嵌入。那么* $L_\theta \le -\mathbb{E}_{x_0\sim p_{\text{data}}}[\log p_\theta(x_0)]$*且*

$$\mathcal{L}_\theta = -\sum_{M\subseteq[L],i\in M} \frac{1}{|M|}\frac{1}{\binom{L}{|M|}} \mathop{\mathbb{E}}_{x_0\sim p_{\text{data}}}[\log p_\theta(x_0^i|x_0[M])], \tag{1}$$

其中 $|M|$ *表示集合* M *的大小，且* $p_\theta(x_i \mid x_0[M])$ *表示第*i*个坐标在* $p_\theta(xt)$ *中的条件概率。*

上述命题的证明见附录E。由于 MDM 损失是所有可能填充掩码M损失的线性组合，损失L_θ的最小化器能够学习解决每一个掩码问题。换言之，最优预测器p_θ是第i个标记在所有掩码M条件下，以$x_0[M]$为条件的后验边缘分布。

另一方面，自回归模型（ARMs）通过学习基于所有先前标记（从x^0到x^{i-1}）来预测i^{th}个标记x^i。这相当于通过屏蔽i到$L-1$的位置来预测x^i。因此，ARMs的训练目标可表示为：

$$\log p_\theta(x_0) = \sum_{i=0}^{L-1} \log p_\theta(x_0^i|x_0[\{i,\dots,L-1\}]). \tag{2}$$

通常，ARMs被训练为从左到右顺序预测标记。我们称之为从左到右的训练。然而，也可以训练这些模型基于序列的固定*已知*排列顺序来顺序预测标记。我们将这种通用方法称为**顺序感知训练**。

为理解MDMs与ARMs训练目标之间的比较，我们希望强调任意阶自回归损失与 MDM 损失的等价性（Hoogeboom 等人，2021a；Ou 等人，2024）。特别地，在命题2.1的条件下， MDM 损失等于

$$\mathcal{L}_\theta = - \mathop{\mathbb{E}}_{\substack{x_0 \sim p_{\text{data}} \\ \pi \sim \text{Unif}(\mathbb{S}_L)}} \left[\sum_{i=0}^{L-1} \log p_\theta \left(x_0^{\pi(i)} \middle| x_0[\pi\{i, \ldots, L-1\}] \right) \right],$$

其中Unif（SL）表示长度L的所有排列的均匀分布（证明详见附录E.1）。值得注意的是，若期望仅针对恒等排列计算，则损失函数将转化为自回归损失。这表明MDM 损失比ARM损失能*按指数级*解决更多子问题。与ARM损失不同， MDM 在训练过程中不偏好任何特定顺序（例如从左到右），因此我们称其训练为*顺序无关训练*。

2.1.2. 无序推断MDMS MDM 推断可分解为两个步骤：

(a) 随机选择一组位置进行解掩码，并(b)通过去噪网络p_θ为每个位置分配标记值。更准确地说，我们可以将逆过程$x_s \sim g_\theta(\cdot|xt)$重新表述如下。

香草 MDM 推断

(a) 采样一组掩码标记$S \subseteq \{i \mid xit = 0\}$，$\mathbb{P}(i \in \mathcal{S}) = \frac{\alpha_s - \alpha_t}{1 - \alpha_t}$.

(b) 对于每个$i \in S$，进行采样$xis \sim p_\theta(x^i | xt)$。

因此， MDM 中的推理是通过随机选择S，然后根据后验概率$p_\theta(xis|xt)$填充每个标记值来实现的。

另一方面，自编码器（ARM）被训练为按从左到右的顺序预测标记，因此生成的标记也遵循从左到右的顺序。相比之下，纯 MDM 推理生成的标记则是随机顺序。

3. MDM在难题上受训

在本节中，我们提供了理论和实证证据，表明当数据分布具有从左到右的顺序（或任何已知固定顺序）时，按从左到右顺序（或已知顺序）进行自回归训练比MDMs更易于处理。特别地，对于此类具有固定顺序的分布，我们证明了ARMs可以高效地从这些分布中采样，而对于MDMs，我们从理论和实证上证明了大部分掩码子问题$p_\theta(xi0|x_0[M])$可能难以学习。

在第3节。1，我们展示了若干简单且非病态的分布示例，对于这些示例：(1)在订单*感知*训练过程中遇到的掩码问题（例如在ARM中）具有可计算性，但(2)许多在无序训练（如MDM）中遇到的文本数据具有计算不可行性。在3.2节中，我们通过实证表明文本数据同样存在顺序感知训练与无序训练的计算复杂度差异，因此MDM需要在复杂度各异的子问题上进行训练（具体取决于顺序/掩码）。在3.3节中，我们通过实证证明训练复杂度的差异会导致**子问题间的性能失衡**：基于此类分布数据训练的MDM在简单子问题上误差较小，但在较难子问题上则会出现较大误差。

3.1. 具有硬掩蔽问题的良性分布

我们现描述一个简单数据模型，基于该模型探讨掩蔽问题的计算复杂度，并展示MDMs与ARMs所遇到的掩蔽问题之间的差异。

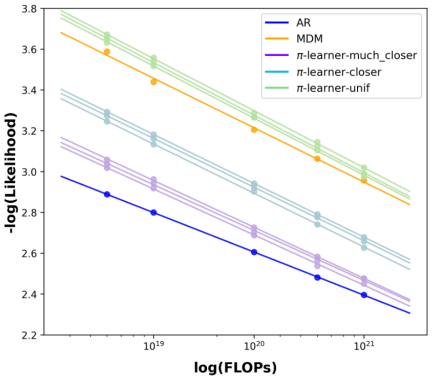
定义 3.1.一个*潜在变量与观测变量（L&O）分布*是长度为L、字母表大小为m（具体而言，p_{data}定义在$\{0, \ldots, m\}^L$上）的数据分布p_{data}，由索引$\{1,2, \ldots, L\}$上的置换π、潜在标记数N、观测标记数P（满足$N+P=L$）以及潜在变量在$\{1, \ldots, m\}$上的先验分布p_{prior}和可高效学习的*观测函数*$\mathcal{O}_1, \ldots, \mathcal{O}_P$：$\{1, \ldots, m\}^N \to \Delta(\{0, \ldots, m\})$，[1]

- （**潜在标记**）对于$i = 1, \ldots, N$，从潜在变量的先验分布p_{prior}中独立采样$x^{\pi(i)}$。
- （**观测标记**）对于$j = 1, \ldots, P$，样本$x^{\pi(N+j)}$独立于$\mathcal{O}_j(x^{\pi(1)}, \ldots, x^{\pi(N)})$。

L&O分布包含两种类型的标记：(1)*潜在标记*和(2)*观测标记*。直观来说，潜在标记是序列中的标记，由$\pi(1)$、$\pi(2)$，...，$\pi(N)$索引，它们作为“种子”为序列提供随机性；其余标记称为观测标记（由$\pi(N+1)$、$\pi(N+2)$，...，$\pi(N+P)$)索引），通过$\mathcal{O}_1, \ldots, \mathcal{O}_P$作为潜在标记的（可能随机化的）函数确定。注意由置换π指定的L&O分布具有通过置换π的自然生成顺序。

订单感知训练订单感知训练，即通过置换序列使得π成为恒等置换，然后进行自回归训练，该方法在计算上是可行的：当$i \le N$时，由于标记是独立的，预测$x^{\pi(i)}$给定$x^{\pi(1)}, \ldots, x^{\pi(i-1)}$是平凡的；当$i > N$时，由于$x^{\pi(i)}$仅

[1]此处*高效可学习*在标准PAC意义上指：给定多项式数量级的形如(z, y)的样本，其中$z \sim p_{\text{prior}}^N$且$y = \mathcal{O}_j(z)$，存在一个高效算法，其在$p_{\text{prior}}^N$上以高概率学习到期望近似$\mathcal{O}_j$。



Task	$\pi = \text{id}$	$\pi \sim \text{Closer}(S_N)$	$\pi \sim \text{Unif}(S_N)$
Val. Loss	3.171	3.212	3.245

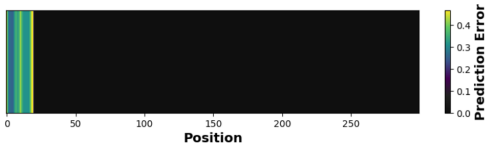


图2. **左图：MDMs在难题上训练（第3.2节）**。x轴和y轴分别对应log（FLOPs）和-log $p_\theta(x)$。MDM（蓝色）在似然建模中表现逊于ARM（橙色）。MDM 训练的大多数掩码问题（其他线条）比ARM遇到的更难，如小的对数似然所示。**右图：任务误差失衡（第3.3节）**。MDM 在不同任务上的表现存在差异。对于文本数据（右上），这通过验证损失体现；对于L&O-NAE-SAT（右下），MDM 在观测位置的掩码问题上表现良好（浅色区域），但在潜在位置上表现不佳（深色区域）。

依赖于$x^{\pi(1)}$、$x^{\pi(N)}$，且根据假设可高效学习。相反，下文将展示示例：若采用与MDMs*类似*的无序训练方法，极有可能遇到严重的掩码问题。

与顺序无关的训练首先我们注意到，如果观测值（$\mathcal{O}_1$，...，OP）由密码学哈希函数生成，那么在给定（$x^{\pi(N+1)}$，...，$x^{\pi(N+P)}$）的情况下预测（$x^{\pi(1)}$，...，$x^{\pi(L)}$）的掩码问题在设计上是计算不可行的，因为它需要逆向哈希函数。虽然这是关于标记顺序在语言建模中作用的众所周知的民间观察，但它并不完全令人满意，因为这种构造本质上是最坏情况——在现实世界的数据中，人们很少在由密码学哈希函数生成的序列上进行训练。此外，它仅针对特定掩码模式建立了难度，而这些模式在运行逆向过程时未必会遇到。

我们提供了几个简单的L&O分布实例来解决这些问题：它们不是利用微妙的密码学构造，而是本质上是*平均情况*的，而且我们可以为逆向过程中遇到的*典型*掩蔽问题建立难度。

在所有这些例子中，即使算法知道所有参数of p_{data}以及观测函数$\mathcal{O}_1, \ldots, \mathcal{O}_P$，我们建立的硬度结果仍然成立。由于篇幅限制，我们在此集中讨论以下例子，将另外两个例子推迟到附录B.1和B.2中讨论。

示例 3.2（稀疏谓词观测）。*考虑以下类别的 L&O 分布。给定项数$k \geq 2$，固定一个谓词函数$g : \{1, \ldots, m\}^k \to \{0, 1\}$。考虑 $\{1, 2, \ldots, N\}$ 的所有有序子集的集合*

将k值设为该集合的大小，并将观测潜变量的总数P设定为该集合的大小（因此$P = N!/(N-k)!) = N(N-1)\cdots(N-k+1)$）。为了采样一个新的序列，我们首先从先验分布$p_{prior}$中采样潜在标记$x^{\pi(1)}, \ldots, x^{\pi(N)}$，并给出与k大小子集S对应的观测潜在变量，其表达式为$g(\{x^{\pi(i)}\}_{i \in S})$。换言之，每个观测潜在变量对应于{1，2，...，N}的k大小子集S，而对应的观测函数$\mathcal{O}_S(x^{\pi(1)}, \ldots, x^{\pi(N)})$由$g(\{x^{\pi(i)}\}_{i \in S})$给出。

命题 3.3. *设x 是来自 L&O 分布p_{data}的样本，该分布具有稀疏谓词观测（如示例3.2所定义），其元数为k，且谓词g 满足假设B.11；设γ表示随机分配{1，...，m}^{k}时g 被满足的概率。设D_{KS}和D_{cond}是与谓词函数g 相关的一些常数（参见定义B.12）。假设x 中的每个标记以概率α被独立屏蔽，M 是屏蔽标记的索引集合。如果$1 - \gamma^{-1} D_{\text{KS}}/kN^{k-1} \leq \alpha \leq 1 - \gamma^{-1} D_{\text{cond}}/kN^{k-1}$，那么在1RSB 空腔预测下（参见猜想B.13），在屏蔽随机性的概率$\Omega_k(1)$下，没有多项式时间算法能够解决给定x[M] 时预测$x^{\pi(1)}, \ldots, x^{\pi(N)}$中任意屏蔽标记的子问题。*

该命题的完整证明见附录B.4。为便于全面理解，我们还在附录B.3中提供了证明概要。

3.2. 基于似然的硬度实证研究

在前一节中，我们提供了理论依据，证明当数据具有自然顺序时，顺序感知训练是可行的，而顺序无关训练则不可行。本节中，我们将通过实证数据支持这一论断。

自然文本数据。此外，近期研究（Nie 等人，2024；Zheng 等人，2024）表明，与自回归模型（ARMs）相比，掩码扩散模型（MDMs）在自然文本数据上的表现较差。本节提供的证据表明，这种性能差距主要源于MDMs的无序训练特性。自然文本本质上遵循从左到右的标记顺序，我们发现当训练偏离该顺序时，模型性能会逐渐下降。

为理解训练过程中顺序的重要性，我们采用以下设定：给定索引{0,1，...，L-1}的排列π，将π *学习器* 定义为如下给出的似然模型$\log p_\theta(x0)$：

$$\log p_\theta(x_0) = \sum_{i=0}^{L-1} \log p_\theta\left(x_0^{\pi(i)} \middle| x_0[\pi\{i, \ldots, L-1\}]\right) \quad (3)$$

换言之，π-学习器在给定干净标记$x\pi_0^{(0)}$、$x\pi_{0(i-1)}$以及掩码标记$x\pi_0^{(i)}$、$x\pi_{0(L-1)}$的情况下，预测位置$\pi(i)$处的标记。若π为恒等置换，则该问题可简化为标准的（从左到右）自回归训练。需注意，MDM损失对每个置换π编码了一个π-学习器，因为MDM损失（1）等同于从Unif（SL）中采样π上所有π-学习器的平均损失：

$$\mathcal{L}_\theta = - \mathop{\mathbb{E}}_{\substack{x_0 \sim p_{\text{data}} \\ \pi \sim \text{Unif}(\mathbb{S}_L)}} \left[\sum_{i=0}^{L-1} \log p_\theta \left(x_0^{\pi(i)} \middle| x_0[\pi\{i, \ldots, L-1\}] \right) \right],$$

其中$\mathbb{S}_L$表示{0,1，...，L-1}上的所有排列集合。上述等价性的证明见附录E。因此，通过测量每个π学习器的“难度”，我们可以探究任意掩码问题与从左到右掩码问题之间的难度差异。

实验装置。我们使用Slimpajama数据集（Soboleva 等人，2023）来评估不同训练顺序的性能。为训练π学习器，我们采用带有因果注意力的Transformer模型，并使用置换数据$\pi(x0)$作为输入。通过在保持所有其他训练配置（如模型、优化）不变的情况下改变π，我们可以将计算得到的似然值（通过公式（3）计算）作为度量指标，用以反映π学习器解决子问题的难度。

在我们的实验中，序列长度L为2048，因此无法对每个π重复缩放定律。相反，我们对π^{-} Unif（SL）进行采样，并考察π学习器似然的缩放定律。我们利用了（Nie 等人，2024）的代码库，其中引入了 MDM 和ARM的基线缩放定律。此外，鉴于RoPE具有

鉴于存在从左到右排序的归纳性偏好，我们在所有实验中采用可学习的位置嵌入层以纠正这一偏差。因此，我们重新运行了采用RoPE的基线结果。为探究π与恒等置换之间的距离如何影响标度律，我们考虑了两种插值分布：一种是均匀分布（SL）（即 MDM 训练），另一种是相同置换下的点质量（即ARM训练）。我们从插值分布和均匀分布（SL）中各抽取三个置换样本，并绘制每个置换对应的标度律曲线。由于篇幅限制，更多实验细节详见附录C.1。

结果。如图2所示，对于具有均匀随机π的π学习器，其缩放定律比ARM更差。这阐明了掩码问题$p_\theta(x_i \mid x_0[M])$在左到右预测之外的固有难度，同时也解释了为何同时在所有$\pi \in \mathbb{S}_L$上训练的 MDM 在似然建模中表现逊于ARM。此外，随着π更接近恒等置换，标度律也更接近ARM（图2中的π-学习器更近和π-学习器更近）。这也支持了ARM适合文本数据的普遍观点，因为它本质上遵循*从左到右*的顺序。

话虽如此，也应注意到，尽管MDM在掩码问题上的训练量比ARM呈指数级增长（$\Theta(L2L)$对比L），但其性能并不显著逊色于π学习器。我们将此归因于*任务多样性的优势*：多任务训练能通过跨任务正向迁移，同时优化优化动态（Kim 等人，2024）和验证性能（Tripuraneni 等人，2021；Maurer 等人，2016；Ruder，2017）。

3.3. 错误在不同掩蔽问题中分布不均

在前几节中，我们已经证明了不同掩码问题的硬度$p_\theta x^{i}|x_0[M])$可能存在显著差异，这可能阻碍 MDM 的学习。本节通过实证数据表明，MDM 最终表现的子问题间失衡程度相似。具体细节参见附录C.2。

L&O-NAE-SAT.考虑一个由π给出的L&O分布，其中每个观测值$\mathcal{O}_j$由NAE（$x_{i_1}, x_{i_2}, x_{i_3}$）确定性地给出，即**1** -[$x_{i_1} = x_{i_2} = x_{i_3}$]，对应某个随机选择的（前缀）三元组（$i_1, i_2$，$i3$）$\in[N]$。对于基于该分布训练的 MDM，我们通过$\mathbb{E}_{x_0}$“$\log p_\theta(x_0|x_0[M])$-来测量其在每个任务对$\log p_\theta(x_0|x_0[M])$的误差。$\log p_{\text{data}}(x_0|x_0[M])\Big\|^2$，其中$p_{\text{dat}}$“$_a(x_0|x_0[M])$表示贝叶斯最优预测器。从技术层面来说，我们无法直接获取该数据，因此改用另一个 MDM 进行大量迭代训练，并将其作为替代指标。图2

结果显示，潜在位置（浅色区域）的预测任务相较于观测位置（深色区域）的预测任务表现出更大的误差。

文本。这里我们重新审视第3.2节的文本实验。由于无法获取贝叶斯最优预测器，我们采用$_{\text{Ex 0 ~ p数据}}$作为度量标准。$\left[\sum_{i=0}^{L-1} \log p_\theta\left(x^{\pi(i)}_0 \mid x_0[\pi\{i, \ldots, L-1\}]\right)\right]$.

这捕捉了子问题$p_\theta\left(x^{\pi(i)}_0 \mid x_0[\pi\{i, \ldots, L-1\}]\right)$中误差的累积，因为$p_\theta(x_0|x_0[M])=p_{\text{data}}(x_0|x_0[M])$最小化了该度量。图2显示了不同子问题之间的明显差距。

理论和实证证据表明，MDMs在估计某些子问题M的$p_\theta(x_0|x_0[M])$时表现优于其他子问题。因此我们希望在推理时避免遇到困难子问题M。在下一节中，我们将证明虽然原始 MDM 推理可能遇到此类子问题，但通过在推理阶段进行简单修改即可有效规避这些问题，从而实现显著的*无需训练*的性能提升。

4. MDM可围绕难题进行规划

我们之前认为，由于掩码子问题的复杂性， MDM 在某些子问题上表现不佳$p_\theta(x^i|xt)$。因此，在标准 MDM 推理过程中， MDM 不可避免地会在步骤(b)遇到这些困难子问题。虽然这可能意味着我们需要从根本上重新审视MDM的训练方式，但本节将展示，令人惊讶的是，在推理阶段进行简单的修改——*而无需任何进一步的训练*——可以避开这些问题，并带来显著的性能改进。

MDM 提供多种采样路径。标准 MDM 推理（算法1）旨在使中间分布与连续扩散中使用的前向过程对齐。然而，与连续扩散不同， MDM 的逆向过程允许多种有效的采样路径（不同顺序的标记解掩码），这些路径与 MDM 前向过程的起始分布相匹配。

我们首先证明，当存在一个完美解决所有掩码问题的理想 MDM 时，即$p_\theta(xi0|x_0[M])=p_{\text{data}}(xi0|x_0[M])$，那么使用任何采样路径（以任意顺序解掩码标记）都会得到相同的分布。考虑以下采样器：对于每一步，S是一个集合，其索引被无意识地选择（不遵循任何分布）。对于由该采样器生成的任何干净样本x_0，注意$p_\theta(x_0)=\prod_{i=0}^{L-1} p_\theta\left(x_0^{\pi(i)}|x_0[\pi\{i,\ldots,L-1\}]\right)$ by chain rule, and这等于$\prod_{i=0}^{L-1}{}_{\text{数据}}\left(x^{\pi(i)}_0 \mid x_0[\pi\{i, \cdots, L-1\}]\right)=$

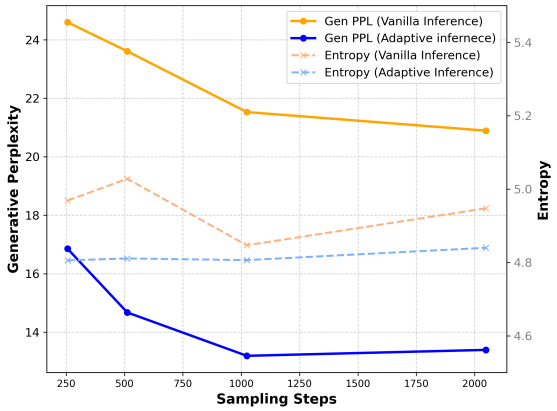


*图3.***生成困惑度.**我们比较了自适应 MDM 推理与传统 MDM 推理的生成困惑度（GenPPL）。我们分别采用预训练的1.7亿 MDM 和LLaMA-7B（Touvron等，2023）作为推理和评估模型。自适应 MDM 推理（蓝色）在保持熵值的同时，显著降低了生成困惑度。

$p_{\text{data}}(x0)$。因此，其他选择的S（不一定遵循算法1）仍能捕捉到真实似然。

在实际应用中，与理想情况不同，如3.3节所示，MDM 在所有子问题上的表现并不均衡。因此，不同的采样路径会导致似然建模能力的差异。基于这一观察，我们考虑对*多变量决策模型进行自适应推断*：

自适应 MDM 推理

(a) 采样一组掩码标记$S=F(\theta, xt)\subseteq\{i \mid xit=0\}$。

(b) 对于每个$i\in S$，样本$xis\sim p_\theta(x^i|xt)$。

自适应 MDM 推断不是随机选择S，而是利用预言机$F(\theta, xt)$来有策略地选择S以避免硬掩码问题。这自然提出了如何设计一个有效的预言机F的问题。

在后续章节中，我们将通过精心选择F参数，证明自适应 MDM 推断能显著提升 MDM 的似然匹配能力。换言之，一个预训练的 MDM，即使它在某些困难的子问题上表现不佳，*仍然包含足够的信息来避免它们*，当与有效的预言机F配对时。

4.1. 订购预言机的有效设计

我们引入两种不同的预测方法：最高概率法和最高概率边际法。直观来说，这两种策略都基于这样的理念：选择S时，应根据模型对每个位置的确定性程度来决定。需要特别说明的是，这些策略不应与核概念等概念相混淆。

在ARMs中的采样（Holtzman 等人，2019）；我们描述的预言机用于选择下一个待解码令牌的*位置*，而非*值*，因此仅在MDMs的上下文中才有意义。

*表1.***L&O-NAE-SAT**。自适应 MDM 推断比普通 MDM 推断实现了更好的似然匹配。注意，朴素猜测的准确率为75%，这表明普通推断的表现与朴素猜测相当或更差。

(N, P)	范式推理	自适应推理
$(25, 275)$	78.06%	93.76%
$(30, 270)$	75.70%	93.54%
$(40, 260)$	74.60%	92.21%
$(50, 250)$	67.94%	90.01%
$(100, 200)$	62.84%	88.91%

最高概率（Zheng 等人，2023）。假设我们希望在时间步t时解掩K个位置，即选择$|S|=K$。在最高概率中，位置的不确定性通过词汇表中任何值的最大概率来估计。更精确地说，位置i的确定性为$\max_{j\in\{0,\ldots,m-1\}} p_\theta(x^i=j|xt)$ 和 $F(\theta, x_t)=\text{Top } K(\max_\theta(xi \mid xt))$.

最高概率策略是许多任务的良好替代方案，在实践中表现良好（Zheng 等人，2023；Ye 等人，2024；Wang 等人，2024）。然而，这种方法常会提供误导性的不确定性估计。例如当 MDM 在两个标记值之间产生混淆时，会赋予它们几乎相等但较高的概率。此时，根据最高概率进行解掩码仍可能选择解掩码该位置，尽管其存在不确定性。为缓解这一问题，我们提出以下替代策略。

最高概率差。在该策略中，位置的不确定性是通过位置i处两个最可能值之间的绝对差来估计的。更准确地说，如果j_1和j_2是根据位置i中$p_\theta(x^i|xt)$在词汇表中最可能的两个值，那么该位置的确定性由$|p_\theta(x^i=j_1|xt)-p_\theta(x^i=j_2|xt)|$和$F(\theta, x_t)=\text{Top } K(|p_\theta(x^i=j_1|xt)-p_\theta(x^i=j_2|xt)|)$给出。当某位置存在多个概率相近的数值时，采用最高概率加差策略能更准确地评估该位置的不确定性；若仅存在单一最优选择值，则最高概率与最高概率加差策略的效果基本一致。

4.2. 自适应 MDM 推理

本节通过实验验证，自适应 MDM 推断能帮助MDMs规避困难子问题，从而提升似然匹配效果。我们首先展示实验结果。

*表2.*数独解题准确率对比。

方法	# 参数	精度
ARM（无排序）ARM（有排序）	42M	9.73% 87.18%
MDM（香草） MDM（最高概率） MDM（最高概率边缘）	6M	6.88% 18.51% 89.49%

在L&O-NAE-SAT和文本数据上，之后转向我们主要的应用领域——逻辑谜题。

L&O-NAE-SAT与文本数据。对于第3.3节定义的L&O-NAE-SAT分布，我们通过测量预测观测标记的准确性来评估自适应推理的有效性。附录中的表1显示，相较于普通推理，该方法有明显改进。针对文本数据集，我们采用标准指标*生成困惑度*进行评估，该指标通过大型语言模型测量生成样本的似然性。同时计算生成样本的熵值，以确保两种推理策略展现出相似的多样性水平。如图3所示，采用自适应推理后生成困惑度显著降低。更多实验细节详见附录D.1。

逻辑谜题。我们研究两种不同类型的逻辑谜题：数独和斑马（爱因斯坦）谜题。直观来看，数独中某些被遮盖的单元格比其他单元格更容易预测，因此我们希望在推理过程中选择这些更容易预测的单元格。我们评估了自适应 MDM 推理在选择此类单元格时相较于传统 MDM 推理的有效性。[2]

为评估推理方法的性能，我们采用正确解题的百分比作为衡量标准。针对两个谜题，我们使用来自Shah 等人（2024）的训练集和测试集。在数独谜题（表2）中，我们发现自适应 MDM 推理（特别是Top概率边际策略）的准确率（89.49%）显著高于传统 MDM 推理（6.88%）。此外，Top概率边际策略的准确率（89.49%）也高于Top概率策略（18.51%）。如第4.1节所述，这是因为当数独中多个竞争值在特定位置的概率相近时（这种情况很常见），Top概率边际策略能更可靠地评估不确定性。对于斑马谜题，如表3所示，我们观察到一致的结果：Top概率（98.5%）和

[2]先前的研究（Ye 等人，2024）报告称，采用Top-K推理的6M MDM 在数独上达到了100%的准确率。鉴于仅采用Top-K的6M MDM 在我们的数据集上仅达到18.51%（表2），这表明（Ye 等人，2024）中的数独数据集比我们的数据集显著更简单。

最高概率置信区间（98.3%）优于标准 MDM 推断（76.9%）。

*表3.*斑马拼图解题准确率对比。

方法	# 参数	精度
ARM（无排序）ARM（有排序）	42M	80.31 % 91.17 %
MDM（香草） MDM（最高概率） MDM（最高概率边缘）	19M	76.9 % 98.5 % 98.3 %

4.3. 利用自适应 MDM 推理在逻辑谜题中引出序列依赖推理路径

在本节中，我们研究自适应 MDM 推理在寻找任务中正确推理/生成顺序的有效性，这些任务中每个序列都有不同的“自然”顺序。为此，我们将比较自适应 MDM 推理与ARM在数独和斑马谜题上的表现。对于这些谜题，生成的自然顺序不仅不同于从左到右，而且是顺序依赖的。对于此类任务，先前的研究表明，如果在训练过程中未提供关于顺序的信息，ARM会遇到困难（Shah 等人，2024；Lehnert 等人，2024）。因此，为获得强基线，我们不仅考虑未使用顺序信息训练的ARM，还考虑了训练数据中每个序列的顺序信息训练的ARM。需注意后者比前者是更强的基线，因为可以期望通过某种方式教会模型推断正确的顺序。一种受监督的教师强制形式（如Shah等人（2024）；Lehnert 等人（2024）中所实施的），消除了以无监督方式确定正确顺序的问题。

我们在表2中比较了数独的ARM和MDM，在表3中比较了斑马谜题。我们观察到，对于这两种谜题，基于最高概率边界的自适应 MDM 推理不仅优于没有排序信息训练的ARM，而且***甚至优于有排序信息训练的ARM***！这表明，使用自适应 MDM 推理的*无监督*方法来寻找正确的顺序并解决此类逻辑谜题，优于使用ARM的*有监督*方法，并且计算密集度显著降低。

4.4. 自然语言任务中的自适应 MDM 推理

为探究不同推理策略对文本基准测试的影响，我们基于Nie 等人（2025）开发的8B MDM 模型LLaDA进行了改进。实验对比了三种推理策略：基础策略、最高概率策略和最高概率差值策略。结果 见表4.

我们发现，两种自适应 MDM 推理策略——最高概率和最高概率差——始终优于传统 MDM 推理。值得注意的是，在 HumanEval-Multiline（填充）、HumanEval-SplitLine（填充）和Math等高难度任务中，最高概率差明显优于最高概率。这是因为当多个标记具有相似概率时（这类情况在这些高难度任务中较为常见），最高概率差能提供更可靠的不确定性估计。这些结果进一步凸显了为各类任务开发新型复杂自适应推理策略的潜力。实验细节详见附录D.3。

4.5. 从易到难的概括

前文已证明，当训练序列与推理序列源自同一分布时，MDM的无序训练结合自适应推理可在逻辑谜题中表现优异。为验证模型是否掌握了正确解题方法并测试自适应推理的鲁棒性，我们还针对数独的更难谜题对MDM进行了测试。

我们保持训练数据集与Shah等人（2024）提出的相同。Shah 等人（2024）通过从Radcliffe（2020）中筛选出可使用7种固定策略且无需回溯搜索的谜题来创建该数据集。我们将Radcliffe（2020）中剩余的谜题作为我们的困难数据集。因此，这些谜题均采用了一种在训练过程中未见过的策略和/或回溯方法来获得正确解。

我们在硬测试集上评估了MDMs和ARMs的准确率，并将结果汇总于表5。数据显示，基于最高概率边际的自适应 MDM 推理策略（49.88%）再次显著优于基于顺序信息训练的ARMs（32.57%）。值得注意的是，尽管两种方法在更具挑战性的测试集上准确率有所下降，但采用自适应推理的MDMs对这种分布偏移表现出更强的鲁棒性。我们认为这源于MDMs需要解决的填充问题数量远超ARMs（exp（L）对比L），因此能比ARMs更高效地提取问题知识。

5. 结论

本研究探讨了标记生成顺序对MDMs训练与推理的影响。我们通过理论与实验证据表明，MDMs在困难掩码问题上进行训练。同时证明自适应推理策略可用于规避这些困难问题。在逻辑谜题中，我们发现这会导致显著的

*表4.*LLaDa 8B模型在编码与数学任务中不同推理策略的性能表现。

方法	人类评估-单项	人源多效病毒	人种分裂	数学	MMLU	ROC 故事
香子兰	31.8%	16.5%	14.2%	28.5%	33.2%	21.23%
最高概率	32.9%	20.8%	18.4%	31.3%	**36.5%**	21.10%
顶部概率边缘	**33.5%**	**25.4%**	**22.3%**	**34.3%**	35.4%	**21.41%**

*表5.*难度高数独解题准确率对比

方法	#Param	精度
ARM（带排序）	42M	32.57 %
MDM（随机）	6M	3.62 %
MDM（最高概率）		9.44 %
MDM（最高概率边缘）		49.88 %

其性能提升不仅超越了基础MDM模型，甚至优于采用教师强制训练（teacher forcing）学习正确解码顺序的ARM模型。未来研究的重要方向是突破相对简单的自适应策略，探索更优的生成顺序，例如本文提出的最高概率排序和最高概率差排序。

致谢。JK感谢KiwhanSong就 MDM 训练展开的讨论。KS和 VK 由美国国家科学基金会人工智能机器学习基础研究所（IFML）资助。KS和 VK 感谢维斯塔GPU集群提供的计算支持，该支持由德克萨斯大学奥斯汀分校生成式人工智能中心（CGAI）和德克萨斯高级计算中心（TACC）提供。KS感谢尼山特·迪卡拉（Nishanth Dikkala）就项目展开的初步讨论。SK特别说明：本研究部分得益于陈-扎克伯格倡议基金会（Chan Zuckerberg Initiative Foundation）设立的坎普纳自然与人工智能研究所（Kempner Institute for the Study of Natural and Artificial Intelligence）的资助，以及美国海军研究办公室（Office of Naval Research）通过项目编号N00014-22-1-2377提供的支持。SC获得哈佛大学院长竞争性优秀奖学金基金（Harvard Dean‘s Competitive Fund for Promising Scholarship）资助，并感谢黄布里斯（Brice Huang）和西丹特·莫汉蒂（Sidhanth Mohanty）就人工种植CSPs的计算统计权衡问题展开的启发性讨论。

影响声明

本文深化了对离散扩散模型的理解，为更广泛的机器学习领域做出了贡献。我们的研究可能产生诸多潜在的社会影响，但在此我们并不认为有必要特别强调其中任何一项。

参考文献

Alaoui，A. E. 和 Gamarnik，D. 对称二元感知器采样解的难度。*arXiv 预印本 arXiv:2407.16627*，2024年。

Alekhnovich，M. 关于平均复杂度与近似复杂度的更多讨论。收录于*第44届 Annual IEEE Computer Science 基础研讨会论文集，2003 年，*第298–307页。IEEE，2003年。

奥宾，B.，珀金斯，W.，和兹德博罗夫áL. 对称二元感知机的存储容量。*《物理A：数学与理论》*期刊，52（29）:294003,2019年。

Austin，J.，Johnson，D. D.，Ho，J.，Tarlow，D.，和 van den Berg，R. 离散状态空间中的结构化去噪扩散模型.*NeruIPS*，2021.

Bavarian，M.，Jun，H.，Tezak，N.，Schulman，J.，McLeavey，C.，Tworek，J.，and Chen，M. 语言模型的高效训练以填补中间空白，2022. URL https://arxiv.org/abs/2207.14255.

Bormashenko，O. 随机转位行走的耦合论证.*arXiv预印本arXiv: 1109.3915*，2011.

张H.、张H.、蒋L.、刘C.和Freeman W. T. Maskgit：Masked生成式图像变换器。*CVPR*，2022年。

陈H.与英L. 离散扩散模型的收敛性分析：通过一致化实现的精确方法
提交。*arXiv 预印本arXiv: 2402.08095*，2024。

陈X、Chi RA、Wang X和Zhou D. 大语言模型推理中的前提顺序问题.*arXiv 预印本 arXiv:2402.08939*，2024.

德塞尔，A.，克扎卡拉，F.，摩尔，C.，与兹德博罗夫á, L.模块化网络随机块模型的渐近分析及其算法应用。*Phys. Rev. E*，84:066106,2011年12月。

Devlin，J. 、 Chang，M. -W.、 Lee，K. 和 Toutanova，K. BERT：深度双向变换器预训练在语言理解中的应用。发表于*2019年北美分会会议论文集*

《计算语言学：人类语言技术》第一卷（长篇与短篇论文集），第4171–4186页，2019年。

Gamarnik，D. 重叠间隙性质：优化随机结构的拓扑障碍。*美国国家科学院院刊*，118（41）:e2108492118,2021.

Golovneva，O.，Allen-Zhu，Z.，Weston，J.，和 Sukhbaatar，S. 逆向训练以解除逆转诅咒.*arXiv预印本 arXiv:2403.13799*，2024.

Gong, S., Agarwal, S., Zhang, Y., Ye, J., Zheng, L., Li, M., An, C., Zhao, P., Bi, W., Han, J., 等人.通过自回归模型的适应性调整实现扩散语言模型的缩放.*arXiv预印本 arXiv:2410.17891*,2024年.

Ho，J.，Jain，A.，和 Abbeel，P. 去噪扩散概率模型.*神经信息处理系统进展*，33:6840–6851,2020.

霍夫曼（J.）、博尔戈多（S.）、门施（A.）、布恰茨卡娅（E.）、蔡（T.）、卢瑟福（E.）、卡萨斯（D. d. L.）、亨德里克斯（L. A.）、韦尔布（J.）、克拉克（A.）等。训练计算最优的大语言模型。*arXiv预印本 arXiv:2203.15556*，2022年。

霍尔茨曼（A. Holtzman）、拜斯（J. Buys）、杜（L. Du）、福布斯（M. Forbes）与崔（Y. Choi）合著的《神经退行性变的奇特案例》一文，发表于*arXiv预印本（arXiv:1904.09751）*，2019年。

Hoogeboom，E.，Gritsenko，A. A.，Bastings，J.，Poole，B.，Berg，R. v. d.，和 Salimans，T. 自回归扩散模型.*arXiv预印本 arXiv:2110.02037*，2021a.

霍格博姆，E.，尼尔森，D.，贾伊尼，P.，福尔é，P. 和 Welling，M. 最大后验概率流与多项式扩散：学习分类分布。*NeurIPS*，2021b.

Kaplan，J.，McCandlish，S.，Henighan，T.，Brown，T. B.，Chess，B.，Child，R.，Gray，S.，Radford，A.，Wu，J.，and Amodei，D. 神经语言模型的标度定律.*arXiv预印本 arXiv:2001.08361*，2020.

Kim，J.，Kwon，S.，Choi，J. Y.，Park，J.，Cho，J.，Lee，J. D.，和 Ryu，E. K. 任务多样性缩短了 icl 平台期。*arXiv 预印本 arXiv:2410.05448*，2024.

Kitouni，O.，Nolte，N. S.，Williams，A.，Rabbat，M.，Bouchacourt，D.，和 Ibrahim，M. 因子分解诅咒：哪些标记项是逆转诅咒的基础以及更多。*神经信息处理系统研究进展*，37卷：112329–112355页，2025年。

Krzakala，F. 和 ZdeborováL. 在随机约束满足问题中隐藏安静解。*物理评论快报*，102（23）:238701,2009年。

Lehnert，L.，Sukhbaatar，S.，Su，D.，Zheng，Q.，McVay，P.，Rabbat，M.，和 Tian，Y. 超越 a* 算法：通过搜索动态引导实现更优的变压器规划。2024年。

廖Y、蒋X、刘Q. 基于概率掩码的语言模型，支持任意词序自回归生成.*计算语言学协会第58届年会论文集*，第263–274页. 计算语言学协会，2020.

刘A.、布罗德里克O.、尼珀特M.和布罗克G.V.离散copula扩散。*arXiv预印本 arXiv:2410.01949*，2024a.

Liu，S.、Mohanty，S. 与 Raghavendra，P.：当信念传播的固定点不稳定时的统计推断。发表于*2021 IEEE第62届 Computer Science 基础 Annual Symposium（FOCS）*，第395–405页。IEEE Computer Society，2022年。

刘，S.，南，J.，坎贝尔，A.，斯塔克，H.，徐，Y.，雅科拉，T.以及GóMez-Bombarelli，R. 在生成时思考：具有计划去噪的离散扩散。*arXiv预印本 arXiv:2410.06264*，2024b.

Loshchilov，I. 和 Hutter，F. 脱耦权重衰减正则化.*arXiv预印本 arXiv:1711.05101*，2017.

Lou，A.，Meng，C.，和 Ermon，S. 通过估计数据分布的比率进行离散扩散建模.*ICML*，2024.

Maurer，A.，Pontil，M.，和 Romera-Paredes，B. 多任务表征学习的优势.*JMLR*，17（81）:1–32,2016.

Montanari，A. 基于随机稀疏观测估计随机变量.*欧洲电信学报*，19(4):385–403,2008.

Nie，S.，Zhu，F.，Du，C.，Pang，T.，Liu，Q.，Zeng，G.，Lin，M.，和 Li，C. 文本掩码扩散模型的扩展研究.*arXiv预印本 arXiv:2410.18514*，2024.

Nie，S.，Zhu，F.，You，Z.，Zhang，X.，Ou，J.，Hu，J.，Zhou，J.，Lin，Y.，Wen，J.-R.，和 Li，C. 大语言扩散模型.*arXiv预印本 arXiv:2502.09992*，2025.

Ou，J.，Nie，S.，Xue，K.，Zhu，F.，Sun，J.，Li，Z.，和 Li，C.您提出的吸收性离散扩散模型，实际上隐式地模拟了原始数据的条件分布。*arXiv预印本 arXiv:2406.03736*，2024年。

Papadopoulos，V.，Wenger，J.，和 Hongler，C. 大型语言模型的时间箭头.*arXiv 预印本 arXiv:2401.17505*，2024.

彭，F. Z.，贝泽梅克，Z.，帕特尔，S.，姚，S.，雷克托布鲁克斯，J.，童，A.，和查特吉，P. 面向掩蔽扩散模型采样的路径规划。*arXiv 预印本 arXiv: 2502.03540*，2025年。

Radcliffe，D. G. 300万道带评分的数独谜题，2020年。网址 https://www.kaggle.com/dsv/1495975。

Rector-Brooks, J., Hasan, M., Peng, Z., Quinn, Z., Liu, C., Mittal, S., Dziri, N., Bronstein, M., Bengio, Y., Chatterjee, P., 等人.通过离散去噪后验预测对掩蔽离散扩散模型进行引导 *.arXiv 预印本 arXiv: 2410.08134*,2024年.

Ruder，S. 深度神经网络中多任务学习的概述*.arXiv 1706.05098*，2017.

Sahoo，S.，Arriola，M.，Schiff，Y.，Gokaslan，A.，Marroquin，E.，Chiu，J.，Rush，A.，and Kuleshov，V. 简单有效的掩码扩散语言模型*.神经信息处理系统进展*，37:130136–130184,2025.

Schiff，Y.，Sahoo，S. S.，Phung，H.，Wang，G.，Boshar，S.，Dalla-torre，H.，de Almeida，B. P.，Rush，A.，Pierrot，T.，和 Kuleshov，V.离散扩散模型的简单引导机制。*arXiv 预印本 arXiv: 2412.10193*，2024年。

Shah，K.，Dikkala，N.，Wang，X.，和 Panigrahy，R. 因果语言建模可激发逻辑谜题的搜索与推理能力。*arXiv预印本 arXiv:2409.10502*，2024.

Shi，J.，Han，K.，Wang，Z.，Doucet，A.，和 Titsias，M. K. 简化与广义掩蔽扩散在离散数据中的应用.*NeurIPS*，2024.

Shih，A.，Sadigh，D.，和 Ermon，S. 以正确方式训练和推断任意阶自回归模型.*NeurIPS*，2022.

索博列娃，D.、阿尔-哈提卜，F.、迈尔斯，R.、斯蒂夫斯，J. R.、赫斯特内斯，J. 和 德伊，N. Slimpajama：Redpajama 的 627b 标记清理与去重版本，2023 年 6 月。

索尔-迪克斯坦（Sohl-Dickstein）、魏斯（Weiss）、马赫斯瓦拉纳坦（Maheswaranathan）与甘古利（Ganguli）. 基于非平衡热力学的深度无监督学习. *ICML*，2015.

Song，Y.、Sohl-Dickstein，J.、Kingma，D. P.、Kumar，A.、Ermon，S.和Poole，B.基于分数的随机微分方程生成建模。*ICLR*，2021年。

Touvron，H.，Martin，L.，Stone，K.，Albert，P.，Almahairi，A.，Babaei，Y.巴什利科夫，N.，巴特拉，S.，巴尔加瓦，P.，博萨莱，S.，比克尔，D.，布莱彻，L.，费雷尔，C. C.，陈，M.、Cucurull、G.、Esiobu、D.、Fernandes、J.、Fu、J.、Fu、W.、Fuller、B.、Gao、C.、Goswami、V.、Goyal、N.、Hartshorn、A.、Hosseini、S.、Hou、R.、Inan、H.、Kardas、M.、Kerkez、V.、Khabsa、M.、Kloumann、I.、Korenev、A.、Koura、P. S.、Lachaux、M.-A.、Lavril、T.、Lee、J.、Liskovich、D.、Lu、Y.、Mao、Y.、Martinet、X.、Mihaylov、T.、Mishra、P.、Molybog、I.、Nie、Y.、Poulton、A.、Reizenstein、J.、Rungta、R.、Saladi、K.、Schelten、A.、Silva、R.、Smith、E. M.、Subramanian、R.、Tan、X. E.、Tang、B.、Taylor、R.、Williams、A.、Kuan、J. X.、Xu、P.、Yan、Z.、Zarov、I.、Zhang、Y.、Fan、A.、Kambadur、M.、Narang、S.、Rodriguez、A.、Stojnic、R.、Edunov、S. 和 Scialom、T. Llama 2：开放基础与微调聊天模型。*arXiv预印本 arXiv: 2307.09288*，2023年。

Tripuraneni，N.，Jin，C.，和 Jordan，M. I. 线性表示的可证明元学习.*ICML*，2021.

Varma，H.，Nagaraj，D.，和 Shanmugam，K. Glauber生成模型：通过二元分类实现的离散扩散模型。*arXiv预印本 arXiv: 2405.17035*，2024.

王X、郑Z、叶F、薛D、黄S和顾Q. 扩散语言模型是多功能蛋白质学习器.*ICML*，2024.

徐M、格夫纳T、克赖斯K、聂W、徐Y、莱斯科维奇J、埃蒙S与瓦赫达特A. 基于能量的扩散语言模型在文本生成中的应用.*arxiv预印本 arXiv: 2410.21357*，2024.

叶杰、高杰、龚松、郑亮、蒋晓、李志和孔亮. 超越自回归：离散扩散在复杂推理与规划中的应用.*arXiv预印本 arXiv: 2410.14157*，2024.

张P.、曾G.、王T.和卢W. Tinyllama：一个开源的小型语言模型。*arXiv 预印本 arXiv: 2401.02385*，2024年。

郑K、陈Y、毛H、刘M-Y、朱J和张Q掩码扩散模型本质上是时间无关的掩码模型，其机制依赖于不精确的分类采样。*arXiv 预印本 arXiv:2409.02908*，2024年。

郑L、袁J、余L与孔L. 一种用于文本生成的重参数化离散扩散模型.*arXiv预印本 arXiv:2302.05737*，2023.

A. 相关工作

离散扩散模型。（连续）扩散模型最初基于具有高斯转移核的连续空间马尔可夫链构建（Sohl-Dickstein 等人，2015；Ho 等人，2020）。随后通过随机微分方程理论将其扩展到连续时间（Song 等人，2021）。类似地，离散扩散模型也从离散空间马尔可夫链中发展而来（Hoogeboom 等人，2021b）。具体而言，（Austin 等人，2021）提出了具有多种类型转移矩阵的D3PM。随后，Lou 等人（2024）提出了 SEDD，其整合了理论与实践上稳健的分数熵目标。此外，Varma 等人（2024）；Liu等人（2024b）提出了创新的建模策略，能够将噪声序列中的标记划分为信号（源自干净数据）或噪声（由前向过程产生）。具体而言，Liu 等人（2024b）利用该方法开发出自适应决策器，可智能判断需要去噪的标记。虽然这与我们讨论的自适应推理策略思路相似，但需要强调的是，该方法专门针对离散扩散过程——这类过程会*打乱*标记值而非对其进行掩蔽处理。

掩码扩散模型。同时，吸收转移核因其优于其他核的性能而成为常用选择。基于此，Sahoo 等人（2025）；Shi 等人（2024）将其框架与连续扩散对齐，形成了一种简单且原理性的训练方案，并称之为*掩码扩散模型*。后续研究探索了 MDM 的各个方面。Gong等人。（2024）通过自回归模型的适应性训练高效训练 MDM，参数规模扩展至70亿。Zheng 等人（2024）将MDMs解释为无序学习器，并基于此提出了一种首次命中采样器。Ye 等人（2024）与Gong 等人（2024）证明 MDM 在推理和规划任务中优于自回归模型，强调了其对下游应用的影响。Nie 等人（2024）研究了 MDM 的扩展规律，而Xu 等人（2024）与Liu 等人（2024a）发现当采样步数较少时难以捕捉坐标依赖性，并提出了额外的建模策略来解决该问题。Schiff 等人（2024）研究了使用 MDM 的条件生成，Rector-Brooks 等人（2024）通过引导方法论应对了控制生成数据分布的挑战。Chen 与 Ying（2024）通过理论分析表明，当评分函数估计准确时，抽样误差较小。

任意顺序推理。尽管语言任务通常具有“从左到右”生成词元的自然顺序，但在规划、推理和组合优化等任务中，词元生成的自然顺序可能与“从左到右”大相径庭。虽然基于自回归的主流语言模型在各类任务中表现优异，但多项研究（Golovneva 等人，2024；Chen 等人，2024；Kitouni 等人，2025）表明这种优异表现与任务的训练顺序相关，因而可能导致其脆弱性。例如，Chen等人（2024）发现，仅数学任务中前提顺序的简单置换就会导致30%的性能下降。这种顺序相关脆弱性的根源在于自回归模型固有的“从左到右”特性。多项研究（Li-ao等人、2020）尝试在自回归框架中解决这一问题。特别是（Papadopoulos 等人，2024）通过比较自然语言中从左到右排序与反向（从右到左）排序的似然性，强调了从左到右排序的重要性。

近年来，离散扩散模型作为自回归模型之外处理离散数据的有前景方法崭露头角。此外，离散扩散模型的阶数无关训练机制在推理过程中开辟了多种采样路径，但其训练阶段也面临若干挑战，因此这类方法似乎能有效激发任何阶次推理能力。Zheng 等人（2023）提出了 MDM 自适应推理策略的不同实现方式，但*为何需要这种自适应推理策略的具体原理仍缺乏深入理解*。本研究从多个维度探讨了基础 MDM 训练的特性，并揭示自适应 MDM 推理如何缓解基础 MDM 训练引发的问题，从而有效激发任何阶次推理能力。

我们还想提及同期开展的相关研究工作Peng 等人（2025）提出了一种替代性自适应推理策略，通过基于BERT模型或去噪器本身选择$F(\boldsymbol{\theta}, x_t)$。具体而言，Peng 等人（2025）利用BERT模型或去噪器获取标记的不确定性，随后采用Top-K来决定解掩码的位置。与他们的工作不同，我们拆解了标记排序对 MDM 训练与 MDM 推理的影响，并更全面地理解了自适应推理的动机与优势。此外，我们的结果表明，在存在多个高概率值时，相较于Top-K边界策略，采用Top-K策略在决定解掩码标记时存在局限性。

超越自回归模型。利用非自回归建模学习自然语言的努力始于BERT（Devlin 等人，2019）。非因果方法可以利用对文本数据表示的理解。

（Chang 等人，2022）采用了类似的方法来学习图像表示。基于这些直觉，（Shih等人，2022；Hoogeboom 等人，2021a）提出了任意顺序建模，该方法允许模型以任意期望顺序生成。Shih 等人（2022）也观察到，任意顺序模型默认需要解决的掩码问题数量比自回归模型呈指数级增长。然而，尽管我们的研究表明，在面对这一具有挑战性的任务多样性时进行学习可使模型在推理阶段受益，他们的研究则试图通过减少需要解决的掩码问题数量来降低训练阶段的复杂性。

B. 第3节的技术细节

符号说明。在本节中，我们使用x^i表示向量x的第i个坐标，$z(j)$表示第j个样本。向量$z(j)$的第i个坐标记为$z(j)^i$。

B.1. 另一个例子：稀疏奇偶校验观测

示例 B.1（含噪声稀疏奇偶校验观测）。*设* $m=2$，$k \in \mathbb{N}$，*且* $N^2 \log N \le P \le N^{0.49k}$。*固定*噪声率$\eta>0$，*以及从集合* $\{0,1\}^N$ *中独立均匀随机采样的字符串* $z(1), \ldots, z(P)$（*该集合包含 k-稀疏字符串*）。*对于每个* $j \in [P]$，*定义* $\mathcal{O}_j(x)$ *为当* $\sum_i x^i z(j)^i$ *为奇数（偶数）时，将质量* $1-\eta$ *分配给* 1（2）*和质量* η *分配给* 2（1）*的分布。注意当* $k=O(1)$ *时，这些观测值均可通过暴力搜索高效获取。*

下面我们将证明，在一定范围的掩码分数下，对于相应的L&O分布，其掩码问题的恒定部分在*带噪声的稀疏学习奇偶性*假设（Alekhnovich，2003）下是计算困难的。形式上我们有：

命题 B.2. *设* $0<\alpha<1$ *为任意绝对常数，且* $\eta=1/\mathrm{poly}(N)$ *足够大。设 x 为来自 L&O 分布* p_{data} *的样本，其奇偶性观测值如示例B.1所定义且存在噪声。假设每个标记以概率* α *被独立屏蔽，M 为被屏蔽标记的索引集合。若* $1-1/N \le \alpha \le 1-1/2N$，*则在稀疏学习奇偶性带噪声（SLPN）假设下（参见定义 B.3），对于 M 的恒定概率，任何多项式时间算法都无法解决给定 x[M] 时，从* $x^{\pi(1)}, \ldots, x^{\pi(N)}$ *中预测任意被屏蔽标记的屏蔽问题。*

我们注意到，将观测数据视为*稀疏*奇偶性并利用*稀疏*学习奇偶性带噪声假设对我们至关重要。若改用*密集*奇偶性并采用*标准*学习奇偶性带噪声（LPN）假设，虽然仍能获得掩码问题的难度，但观测数据本身将难以学习（基于LPN假设）。该结论基于以下标准难度假设：

定义 B.3（含噪声的稀疏学习奇偶性）。给定输入维度N、噪声参数$0<\eta<1/2$以及样本量P，*含噪声的稀疏学习奇偶性（SLPN）*问题的实例生成如下：

- 自然从$\{0,1\}^N$中随机采样一个比特串x
- 我们观察到P个形式为（$x(i)$，$y(i)$）的示例，其中$x(i)$是从$\{0,1\}^N$中的k稀疏比特串中独立且均匀随机采样的，而y由$\epsilon_i + Ax(i)$，x（模2）给出，其中ϵ_i以η的概率为1，否则为0。

给定示例 $\{(x(i), y(i))\}_{i=1}^P$, the goal is to recover x.

*SLPN 假设*是对于任意$P=N^{(1-\rho)k/2}$（其中常数$0<\rho<1$），以及任意足够大的逆多项式噪声率η，不存在多项式（N）时间算法能够以高概率恢复x。

命题证明B.2。以概率至少为$1-(1-1/N)^N \ge \Omega(1)$，所有变量标记$x^{\pi(i)}$对于$i \le N$都被屏蔽。独立地，观察标记$\mathcal{O}_j$中未屏蔽标记的数量服从 Bin（$P$，$1-\alpha$）分布，因此根据Chernoff界，以概率至少为$1-e^{-\langle sp_0|\rangle (P/N2)} = 1-1/$多项式（$N$），我们得到至少有$P/4N = \Omega(N \log N)$个观察标记未被屏蔽。在这种情况下，屏蔽问题相当于输入维度为N、样本量为[<sp_0|>（$N \log N$），$O(N^{0.49k})$] 的 SLPN 实例。由于样本量的下界，对$\mathbf{x}^M$的预测在信息论上是可能的。由于样本量的上界， SLPN 假设使得计算上变得困难。因此，在给定未掩码标记的情况下，对$\mathbf{x}^M$的任意条目估计后验均值是

如权利要求所述的计算复杂度高 □

B.2. 另一个例子：随机平板观测

示例 B.4（随机板面观测）。*设 $m=2$ 且 $P=\gamma N^2$ 对于常数 $\gamma>0$。固定板面宽度 β 以及从 $N(0,I)$ 中独立采样的向量 $z(1),\ldots,z(P)$。对于每个 $j\in[P]$，定义相应的观测值 $\mathcal{O}_j(x)$ 为：若 $|x(j)|\le\beta\sqrt{N}$ 则确定性地 1，否则确定性地 0。*

在（Alaoui & Gamarnik，2024）中，研究表明*稳定*算法（定义B。7），其中包含许多强大的统计推断方法，如低次多项式估计器、 MCMC 和算法随机定位（Gamarnik，2021），但无法从后验分布中采样出满足 $|x(j),x\rangle|\le\beta\sqrt{N}$ 的随机比特串，对于任意 $\Theta(N)$ 个约束条件 $z(1),\ldots,z(P')$，前提是 P' 不过大到使后验分布的支持集为空。该集成是研究充分的*对称感知器*（Aubin 等人，2019）。以下是对（Alaoui & Gamarnik，2024）结果的直接重新解释：

命题 B.5. *Let p_{data} 是一个具有随机板状观测的L&O分布，如示例B.4所定义，其参数为 $\gamma>0$，板宽 $\beta>0$。存在一个常数 $c_\beta>0$，使得对于任意绝对常数 $0<c<c_\beta$，若 $1-c_\beta N/2P\le\alpha\le 1-cN/P$ 且 $\gamma>c_\beta$，则以下条件成立。设 p'_{data} 表示通过以概率 α 独立掩码 p_{data} 中每个坐标所得到的分布。那么任何（$1-\Omega(1/\sqrt{N})$）-稳定的算法，即使不基于掩码扩散的算法，只要它以概率 $1-o(1)$ 从 p'_{data} 中获取样本 x'，并以概率从 x' 中未掩码的标记条件输出一个Wasserstein近似[3]样本 p_{data}，该算法必须在超多项式时间内运行。*

其最终结果是，任何稳定的多项式时间掩码扩散采样器，在逆向过程的某个阶段，将以不可忽略的概率遇到计算上困难的掩码问题。

在证明过程中，我们首先正式定义（植入式）对称伊辛感知机模型：

定义 B.6. 设 $\alpha,\beta>0$。*种植对称Ising感知器*模型定义如下：

- 自然样本 σ 均匀随机地从 $\{\pm1\}^N$ 中
- 对于每个 $j=1,\ldots,P=\lfloor\alpha N\rfloor$，我们从 $N(0,IN)$ 中独立采样 $z(j)$，条件是满足 $|x(j),\sigma\rangle|\le\beta\sqrt{N}$。

目标是从后验分布中采样，σ 条件于这些观测值 $\{z(i)\}_{i=1}^P$.

接下来，我们将形式化*稳定算法*的概念。

定义 B.7. 给定一个矩阵 $Z\sim N(0,1)^{\otimes P\times N}$，定义 $Z_t=tZ+\sqrt{1-t^2}Z'$，其中 $Z'\sim N(0,1)^{\otimes P\times N}$ 是独立的。一个随机算法 A，其以 $Z\in\mathbb{R}^{P\times N}$ 作为输入，并输出一个 $\{\pm1\}^N$ 的元素，若满足 $\lim_{N\to\infty}W_2(\text{law}(A(Z)),\text{law}(A(Z_t)))=0$，则称其为 t_N-*稳定*。

正如Gamarnik（2021）中深入讨论的，许多算法如低度多项式估计器和朗之万动力学是稳定的。

定理 B.8（（Alaoui & Gamarnik，2024）[4]）*中定理 2.1）。对于任意常数 $\beta>0$，存在 $c_\beta>0$，使得以下条件对所有常数 $0<\alpha<c_\beta$ 成立。对于 $t_N\le 1-\Omega(\log^2(n)/n^2)$，任何以 $Z=(z(1),\ldots$ 为输入的 t_N-稳定随机算法 A，$z(P))$ 并输出一个 $\{\pm1\}^N$ 的元素将无法从对称Ising感知器模型中 σ Z 条件下的后验分布中采样，其Wasserstein误差为 $o(\sqrt{N})$。*

*命题B.5*的证明。通过联合界，对于一次抽取 $x'\sim p'_{\text{data}}$，所有 $x^{\pi(i)}$ 令牌被掩码的概率至少为 $1-(1-\alpha)N\ge 1-c_\beta N^2/P\ge 1-c_\beta/\gamma$。在 x' 中未被掩码的令牌数量在观测值 $\mathcal{O}_j$ 中服从 $\text{Bin}(P,1-\alpha)$ 分布。根据切尔诺夫界，该概率至少以常数概率落在 $[3cN/4,3c_\beta N/4]$ 区间内。因此该结论可直接由上文定理B.8得出。 □

B.3. 命题3.3的证明大纲

为理解该证明思路，我们考虑所有潜在标记均被掩码且部分观测标记未被掩码的情况。在此情形下，预测任务可简化为学习恢复与观测数据一致的潜在标记。

[3]此处的近似概念是Wasserstein-2距离中的 $o(1)$-接近性。

[4]需要注意的是，虽然Alaoui & Gamarnik（2024）中的定理陈述指的是对称二元感知器的非种植版本，但他们证明的第一步是论证这两个模型在感兴趣的领域中是相互连续的。

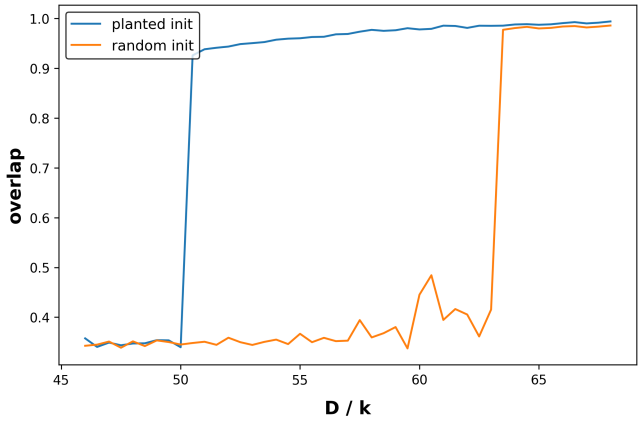


图4. 与随机相比，基于真实值初始化的信念传播在种植CSP中实现的重叠度，其中$k=3$，$m=3$，且$g=\mathrm{NAE}$，对于$N=10000$及平均度D的不同选择。D_{KS}/K可通过解析证明为64，与所描绘的相变一致。图示表明$D_{\mathrm{cond}}/K \approx 50$。根据命题3.3，这意味着存在一系列掩蔽分数，在此范围内$\boldsymbol{\Omega}(1)$比例的掩蔽问题具有计算难度。

根据观测数据，直观来看，每个观测都会带来一定约束条件，而任务就是找出满足这些约束的分配方案。这让人联想到*约束满足问题*（CSPs）。实际上，为了证明该问题的难度，我们采用了统计物理与平均复杂度理论交叉领域中为CSPs研究而发展出的丰富理论。

在植入式CSP中，存在一个长度为N的未知随机采样向量y，并给定随机选择的布尔约束条件，这些约束条件承诺y必须满足，目标是尽可能准确地恢复y（参见定义B.9）。先前的研究表明，高效学习解决植入式CSP问题的难度（Krzakala & Zdeborová，2009；Alaoui与Gamarnik，2024）。我们基于这些结果展示了在L&O分布中掩码问题的难度。将真实潜在标记视为随机向量y，并将每个观测视为约束条件。在这种情况下，从观测标记中学习恢复潜在标记的问题可简化为对植入CSP的恢复。

对于词汇量m和观测值数量的精确预测表明，信息论上最佳可能重叠与任何计算高效算法可实现的最佳重叠存在差异。我们证明这些预测可直接转化为关于掩蔽问题何时变得计算不可行的预测：

作为一个简单的例子，让我们考虑稀疏谓词观测，其中$k=2$且$g(x', x'') = \mathbf{1}[x' \neq x'']$。这些可以形式化地与研究充分的*植入$m$-着色*问题相关联。在植入$m$-着色中，采样一个平均度为$D$的随机图，该图与未知顶点着色一致，目标是尽可能准确地估计着色（Krzakala & Zdeborová，2009），通过算法输出与真实着色*重叠*度量（参见定义B.9）。作为我们主要结果的推论，我们证明当所有潜在标记$x^{\pi(1)}, \ldots, x^{\pi(N)}$被掩码且少数未掩码的观测标记提供形式为$g(x^{\pi(i)}, x^{\pi(j)}) = \mathbf{1}[x^{\pi(i)}$的信息时$\neq x^{\pi(j)}]$对于$i, j \leq N$，则可将掩蔽问题简化为求解种植着色问题。

对于有向图m着色，当$m=5$时，命题3.3中的阈值由$D_{\mathrm{KS}}/2 = 16$和$D_{\mathrm{cond}}/2 \approx 13.23$给出（Krzakala & Zdeborová 2009）（此处的2倍因子仅是因为观测值对应于大小为2的*有序*子集）。对于一般谓词和算子，基于*信念传播*算法的行为，已有成熟的数值计算D_{KS}和D_{cond}的方法（参见附录B中的讨论。4）。例如，在图4中，我们对$m=3$、$k=3$以及由非全等谓词NAE给出的$g(x', x'', x'') = 1\text{-}\mathbf{1}[x'=x''=x''']$执行该方案，以获得可代入命题3.3的阈值。

硬度的更多示例。上述设置还可推广以捕捉*贝叶斯约束满足问题*（Montanari，2008；Liu 等人，2022），其中典型范例是随机块模型（Decelle 等人，2011）。对于推理难度的起始存在类似预测，这些预测同样可转化为看似良性L&O分布的掩蔽问题的难度。详见附录B.1和B.2，我们给出了两个L&O分布的更多示例，对于这些示例，顺序感知训练是可行的，但 MDM 的顺序无关训练在计算上

困难的

首先，我们考虑观测值稀疏、潜变量存在噪声的L&O分布，并基于带噪声的 Sparse Learning 对称性假设（Alekhnovich，2003）推导出无序训练的难度。接着，我们研究观测值为潜变量*广义线性模型*的L&O分布，并利用对称二元感知器（Aubin 等人，2019）的Lipschitz难度现有结果（Alaoui & Gamarnik，2024），推导出一大类高效算法的难度。

B.4. 命题3.3的证明：稀疏谓词观测

本文将对命题3.3中关于难度的主张进行形式化定义，为此需要明确相关概念。

定义 B.9（种植型 CSP）。给定元数$k \in \mathrm{N}$，词汇表/字母表大小$m \in \mathrm{N}$，谓词g：$\{1, \ldots, m\}^k \to \{0,1\}$，潜在维度$N$，以及子句密度$P/N$，相应的*种植约束满足问题*定义如下：自然从$\{1, \ldots, m\}^N$中均匀随机采样一个未知赋值$\boldsymbol{\sigma}$，然后对于$[N]$中每个有序的$k$-元组$S$，若$g(\boldsymbol{\sigma}|_S) = 1$，则以概率$\boldsymbol{\phi}/N^{k-1}$独立观测到*子句*$S$。

为衡量给定观测值时恢复$\boldsymbol{\sigma}$的算法质量，定义估计$\hat{\boldsymbol{\sigma}}$与真实$\boldsymbol{\sigma}$之间的*重叠度*$d(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}})$。$\min_{\pi \in SN} \sum_i \mathbf{1}[\sigma_i = \pi(\hat{\sigma}_i)]$ 其中$\mathbb{S}_N$表示$\{0,1, \ldots, N-1\}$的所有排列集合。定义*平均度*为kP/N，即与给定变量共享至少一个子句的预期变量数量。

我们首先定义驱动随机约束满足问题硬度统计物理预测的核心算法：置信传播（BP）。

定义 B.10（BP 更新规则）。信念传播是一种迭代更新一组*消息* $\{\mathrm{MSic}^{\to S}[t], \mathrm{MSSc}^{\to i}[t]\}$的算法，其中$i$，$S$覆盖所有变量索引对$i \in [N]$ 与观测值S，i。在时间$t+1$，消息通过

$$\mathrm{MS}_c^{i\to S}[t+1] \propto \prod_{T: i\in T\neq S} \mathrm{MS}_c^{T\to i}[t] \tag{4}$$

$$\mathrm{MS}_c^{S\to i}[t+1] \propto \sum_{\overline{\sigma}\in\{1,\ldots,m\}^{S\setminus i}} g(\overline{\sigma} \cup_i c) \prod_{j: i\neq j\in S} \mathrm{MS}_{\overline{\sigma}_j}^{j\to S}[t], \tag{5}$$

其中$\overline{\boldsymbol{\sigma}} \cup_i c \in \{1, \ldots, m\}^S$将$c$分配给条目$i$，并将$\overline{\boldsymbol{\sigma}}$分配给其余条目。

一组消息可用于估计后验分布的边缘概率，条件是$\boldsymbol{\sigma}$，具体如下。第i个变量的边缘概率具有概率质量函数，其取值范围为$\{1, \ldots m\}$与 $\{\boldsymbol{\Pi}_{T: i\in T} \mathrm{MSTc}^{\to i}\}$ 成正比。给定一组边缘分布，提取$\boldsymbol{\sigma}$估计值的自然方法是将概率质量函数取值四舍五入到 $\{1, \ldots, m\}$中概率最大时的颜色。

在本文中，我们将做出以下假设，以确保平凡消息$\mathrm{MSic}^{\to S}=1/m$和$\mathrm{MSc}^{\to i}=1/m$是上述迭代的固定点，有时称为顺磁固定点：

假设 B.11.量$\boldsymbol{\Sigma}_{\overline{\sigma}\in\{1,\ldots,m\}^{[k]\setminus i}} g(\overline{\boldsymbol{\sigma}} \cup_i c)$在所有$c \in \{1, \ldots, m\}$和$i \in [k]$ 上保持恒定。

定义 B.12.给定k，m，g，*Kesten-Stigum*阈值D_{KS}被定义为BP算子在顺磁固定点附近局部稳定的最大平均度，即从顺磁固定点的小扰动出发，它会收敛到顺磁固定点。更正式地说，D_{KS}是BP算子雅可比矩阵$\{\mathrm{MS}^{i\to S}[t]\} I \to \{\mathrm{MS}^{i\to S}[t+1]\}$ 的最大平均度，其谱半径小于1。

*凝聚*阈值D_{cond}被定义为：当$N\to\infty$时，种植的CSP集成与随后的简单*零模型*在最大平均度数上相互连续且统计上无法区分的临界值。零模型定义如下：不存在单一未知分配，而是对于k个变量的每个有序子集S，自然界独立采样一个未知的局部分配$\boldsymbol{\sigma}_S \in \{1, \ldots, m\}^S$，且若$g(\boldsymbol{\sigma}_S) = 1$，则观测值以概率$\boldsymbol{\phi}/N^{k-1}$被包含。

对于$D_{\mathrm{cond}}<kP/N<D_{\mathrm{KS}}$，存在某个BP算子的*其他*不动点，其边缘分布一旦被四舍五入为一个分配，其重叠度会严格高于随机初始化消息的BP。预测表明，在此情形下，任何高效算法都无法实现最优恢复（Krzakala & Zdeborová, 2009).

推测 B.13（1RSB 空腔预测）。*假设 k、m、g 满足假设 B.11，并令 D_{KS} 和 D_{cond} 分别表示平均度的 Kesten-Stigum 阈值和凝聚阈值。那么对于所有满足 $D_{\mathrm{cond}} < kP/N < D_{\mathrm{KS}}$ 的 P，计算高效算法在恢复 $\boldsymbol{\sigma}$ 时所能达到的最佳重叠度严格小于可实现的最佳重叠度。*

命题证明 3.3。在满足命题中界限的掩码分数 $\boldsymbol{\alpha}$ 下，以至少 $\boldsymbol{\alpha}^N \geq (1-\boldsymbol{\gamma}^{-1} D_{\mathrm{KS}}/Nk\text{-}1)^N \geq \boldsymbol{\Omega}(1)$ 的概率，我们得到所有对应于潜在变量 $x_{\pi(i)}$ 的标记都被掩码。独立于此，观测标记 $\mathcal{O}_S$ 中未被掩码的标记数量服从 Bin（N（$N-1$）⋯（$N-k+1$），$1-\boldsymbol{\alpha}$）分布，因此根据标准二项式尾部界，以常概率（取决于 D_{cond} 与 D_{KS} 之间的差距）该值介于 $\boldsymbol{\gamma}^{-1} D_{\mathrm{cond}}\, N/k$ 与 $\boldsymbol{\gamma}^{-1} D_{\mathrm{KS}}\, N/k$ 之间。此外，这些未被掩码的标记中，期望有 $\boldsymbol{\gamma}$ 比例对应于相关谓词评估为1的观测值。在上述事件条件下，掩码问题因此精确归约为平均度为 $D_{\mathrm{cond}} < D < D_{\mathrm{KS}}$ 的植入约束满足问题的推理，由此
命题随之成立。 □

C. 实验细节见第3节

C.1. 第3.2节中的实验细节

$\boldsymbol{\pi}$-学习器配置。我们考虑两种 $\boldsymbol{\pi}$ 的分布，它们在均匀分布（SL）与点质量分布之间进行插值：其中 $\mathbb{S}_L$ 表示索引 $\{0,1,\ldots,L-1\}$ 所有排列的均匀分布，而点质量分布则为相同分布（更接近）和（更接近）。为构建这些分布，我们从恒等排列开始，执行一定数量的随机交换操作。由于 $L\log(L)$ 次交换会产生一个非常接近均匀分布（SL）的分布（Bormashenko，2011），我们分别使用 $L/10$ 和 $\sqrt{L}$ 次交换来构建（更接近）和（更接近）分布。为保持一致性，我们重复此采样过程三次。

模型与训练配置。如第3.2节所述，为评估 $\boldsymbol{\pi}$ 学习器的扩展规律，我们只需对自回归训练架构（采用因果注意力机制的Transformer）进行调整：将输入修改为 $\boldsymbol{\pi}(x0)$，并用可学习的位置嵌入层替代RoPE。我们借鉴了Nie 等人（2024）的训练配置，这些配置也与TinyLlama（Zhang 等人，2024）的配置一致。具体而言，我们使用AdamW优化器（Loshchilov & Hutter，2017），设置 $\boldsymbol{\beta}_1 = 0.9$、$\boldsymbol{\beta}_2 = 0.95$，权重衰减为0.1且 $L = 2048$。采用余弦学习率调度策略，最大学习率为 4×10^{-4}，最小学习率为 4×10^{-5}。还需注意**除非另有说明，本文始终采用相同的训练配置。**

研究缩放定律。我们进行IsoFLOP分析（Hoffmann 等人，2022）。对于给定的浮点运算次数 C，通过调整Transformer的非嵌入参数数量，我们设定迭代次数，使得模型在训练期间观察到的总标记数等于 $C/6N$，这遵循先前的研究（Hoffmann 等人，2022；Kaplan 等人，2020）。随后我们选择最小的验证损失并将其设为数据点。

C.2. 第3.3节中的实验细节

C.2.1.实验ON L&O-NAE-SAT分布

我们考虑L&O-NAE-SAT分布（N，P）=（20,280）。对于L&O-NAE-SAT中的每个示例序列，我们在最后212个标记后添加一个值为2的额外标记。我们采用19M MDM 并使用RoPE，最大序列长度为512。随后，该 MDM 经过 2×10^3 次迭代训练。为了获得贝叶斯最优预测器的代理 MDM，我们进一步对其进行 5×10^4 次迭代训练。

为测量不同任务间的误差，我们采用以下设置：对于每个 $\ell \in [1, N-1]$，我们在潜在位置随机掩码 ℓ 个标记，在观测位置随机掩码 $\ell\times(P/N)$ 个标记。在所有被掩码的预测位置（$\ell(1+P/N)$）上，我们测量每个位置的误差。为确保可靠性，我们重复此过程1000次。图2中的结果对应 $\ell = 11$ 的情况，其他 ℓ 值也观察到相同趋势。

C.2.2.实验文本数据

我们采用一个基于文本数据预训练的170M MDM 作为基线模型，用以衡量似然建模任务之间的性能差异。

$$\mathbb{E}_{x_0\sim p_{\text{data}}}\left[\sum_{i=0}^{L-1}\log p_\theta\left(x_0^{\pi(i)}\middle|x_0[\pi\{i,\ldots,L-1\}]\right)\right].$$

如第3.2节实验所示，我们从三个不同分布中采样$\boldsymbol{\pi}$：Unif（SL）、（Closer）以及相同分布的点质量。针对每种情况，我们计算了$x_0\sim p_{\text{data}}$的1024个样本的期望值。

D. 实验细节见第4节

D.1. 第4.2节中的实验细节

D.1.1.实验ON L&O-NAE-SAT分布

我们研究了L&O-NAE-SAT的五个实例：（N，P）=（25,275）、（30,270）、（40,260）、（50,250）、（100,200）。针对每个分布，我们训练了一个1900万 MDM，并通过最高概率差值来衡量标准推理与自适应推理之间的准确度差异。

D.1.2.实验文本数据

带温度的最高概率边缘采样器。为了改进我们对文本数据建模的推理，我们发现给预言机添加一定水平的温度是有用的。这是因为最高概率边缘或最高概率通常会导致贪婪采样，这会损害生成样本的多样性（熵）。因此，我们考虑了预言机的变体，如下所示，加入高斯噪声项$\boldsymbol{\epsilon}$。

$$F(\boldsymbol{\theta},\ x_t)=\text{Top}\,K\left(|p_\theta(x^i=j_1|xt)-p_\theta(x^i=j_2|xt)|+\boldsymbol{\epsilon}\right).$$

注意，这种方法也被用于无条件采样（Wang 等人，2024；Zheng 等人，2023）。

生成困惑度与熵。我们采用基于文本数据预训练的1.1B MDM 作为基线。在每个采样步骤中，我们同时使用纯推理和自适应推理无条件生成样本。随后，我们以LLama2-7B作为基线大语言模型计算似然值。此外，我们将生成样本x的熵表示为$\boldsymbol{\Sigma\pi}\log p_i$，其中$p_i=\#\{x^i=i\}/L$。

选择要解掩码的标记数量。我们将未掩码标记的数量设置为K，使得未掩码标记的数量与标准 MDM 推理的期望值相匹配。对于从步骤t到s的推理转换，标准 MDM 期望（当前xt中的掩码标记数量）$\times\left(1-\frac{\alpha_s}{\alpha_t}\right)$未掩码。因此，我们选择$K$=（当前$xt$中的掩码标记数量）$\times\left(1-\frac{\alpha_s}{\alpha_t}\right)$。这一选择在整个推理过程中保持了揭示标记数量的平衡。或者，也可以从Binomial（当前x_t、$\left(1-\frac{\alpha_s}{\alpha_t}\right)$中的掩码标记数量）中随机采样$K$。我们发现，$K$的确定性选择和随机选择都能产生可比的生成困惑度。

该选择K在网络具有时间条件性时可能具有潜在帮助，因为这保持了（当前xt中的）（#mask 个标记）$\approx(1-\boldsymbol{\alpha}t)\times L$，其中$L$是最大序列长度——与模型在训练期间观察到的边缘相匹配。

D.2. 数独与斑马谜题的实验细节

数据集。对于数独和斑马谜题，我们使用Shah 等人（2024）提供的数据集来训练模型。为评估模型在相同难度任务上的表现，我们采用Shah 等人（2024）提出的测试数据集。该数据集通过筛选Radcliffe（2020）中可使用固定7种策略解决的谜题生成。为创建评估易到难泛化的困难数据集，我们采用Radcliffe（2020）中剩余的谜题，因其需要训练期间未见过的新策略和/或回溯操作。该困难数据集包含约100万道数独谜题。

模型、训练与推理。在训练和推理阶段，我们采用（Ye 等人，2024）的代码库，保持其中大部分超参数默认设置。对于数独数据集，我们使用600万参数的GPT-2模型；对于斑马数据集，则采用1900万参数的模型。训练时将学习率设为0.001，批量大小为128，训练300个周期。推理阶段采用50次反向采样步骤并配合相应策略。此外，我们在 MDM 推理预言机F中添加了系数为0.5的Gumbel噪声。

D.3. LLaDA-8B的实验细节

我们的评估涵盖两大任务类别：(i) 填补任务（HumanEval-Infill和ROCStories）和（ii）指导-回答任务（数学）。在指导-回答任务中，我们采用半自回归采样策略；而在填补任务中则保持非自回归方法。填补任务的输出长度是预设的——与被遮蔽片段的大小相匹配——而指导-回答任务需要明确的长度规范。对于后者，我们遵循Nie 等人（2025）的采样配置。

对于HumanEval-Infill，我们采用由（Bavarian 等人，2022）引入的问题集。每个实例根据被遮蔽代码的跨度——即模型必须填补的区域——分为三类：*单行*、*多行*和*分割*。随着被遮蔽跨度的增加，任务难度也随之上升。

E. 遗漏的证明

*命题 2.1*的证明。我们基于命题3进行构建。1 来自（Zheng 等，2024）以获得命题2.1的结果。我们首先重述（Zheng 等，2024）中关于去噪网络p_θ不显式依赖噪声尺度t的结果。设x（n）为从x_0中屏蔽n个标记的序列，且x^i（n）表示序列x（n）的第i个标记值。设
$\tilde{q}(x(n)|x_0)$ be the probability distribution corresponding to randomly and uniformly masking nx_0的代币。

命题 E.1（（Zheng 等，2024）*中的命题 3.1*）。*对于干净数据*x_0, *let* $\tilde{q}(x(n) \mid x0)$*是一个离散前向过程，它随机且均匀地屏蔽*x_0*的*n *个标记。假设噪声调度*α_t*满足*α_0=0*且*α_1=1*。那么， MDM 训练损失（*1*）可以重新表述为*

$$\mathcal{L}_\theta = -\sum_{n=1}^{L} \mathop{\mathbb{E}}_{x(n)\sim\tilde{q}(\cdot|x_0)} \left[\frac{1}{n} \sum_{\ell: x^\ell(n)=0} \log p_\theta(x_0^\ell \mid x(n)) \right]. \tag{6}$$

为获得式（6）的替代表达式，我们展开期望值x（n）$\bar{}$$\tilde{q}(\cdot \mid x0)$。由于存在总共$L$个$x_0$位置，每个$x$（$n$）的分配概率等于1/（$_{Ln}$）。因此，通过期望$x$（$n$）展开上述方程，并将$x$（$n$）视为某个大小为$n$的集合$M$中的$x[M]$，我们得到该结果。

$$\mathcal{L}_\theta = -\sum_{M\in[L], i\in M} \frac{1}{\binom{L}{|M|}} \cdot \frac{1}{|M|} \log p_\theta(x_0^\ell \mid x[M]).$$

□

E.1. MDM 损失与任意阶自回归损失的等价性

本节将论证 MDM 损失与任意阶自回归损失的等价性。具体而言，针对所有x_0，我们证明

$$-\mathop{\mathbb{E}}_{\pi\sim\mathrm{Unif}(\mathbb{S}_L)} \left[\sum_{j=0}^{L-1} \log p_\theta \left(x_0^{\pi(j)} \middle| x_0[\pi\{j\}, \ldots, \pi\{L-1\}] \right) \right] = -\sum_{M\subseteq[L], i\in M} \frac{1}{\binom{L}{|M|}} \frac{1}{|M|} \log p_\theta(x_0^i | x_0[M]).$$

我们现在考虑{π（j），...π（L－1）}=M⊆[L] 且π（j）=i，并统计$\pi \in \mathbb{S}_L$中能诱导特定项 log p $_\theta$（$xi0|x_0[M]$）的数量。要诱导该项，对于给定的$M \in [L]$ 和$i \in M$，π必须满足

π（j）=i，{π (j)，...，π（L－1）}=M。

满足上述条件的$\boldsymbol{\pi}$个数是（L - $|M|$）！×（$|M|$ -1）！. 利用这一点和总排列数为L！，我们得到结果。

$$\mathop{\mathbb{E}}_{\pi\sim\mathrm{Unif}(\mathbb{S}_L)}\left[\sum_{j=0}^{L-1}\log p_\theta\left(x_0^{\pi(j)}\middle|x_0[\pi\{j\},\ldots,\pi\{L-1\}]\right)\right]$$

$$=\frac{1}{L!}\sum_{\pi\in\mathrm{Unif}(\mathbb{S}_L)}\sum_{j=0}^{L-1}\log p_\theta\left(x_0^{\pi(j)}\middle|x_0[\pi\{j\},\ldots,\pi\{L-1\}]\right)$$

$$=\frac{1}{L!}\sum_{M\in[L],i\in M}\left[\log p_\theta(x_0^i|x_0[M])\times(L-1-|M|)!\times(|M|-1)!\right]$$

$$=\sum_{M\in[L],i\in M}\frac{1}{\binom{L}{|M|}}\frac{1}{|M|}\log p_\theta(x_0^i|x_0[M]).$$