arXiv:2602.02927v1 [stat.ML] 2 Feb 2026

多模态掩码扩散模型的免训练自我纠正

Yidong Ouyang[1], Panwen Hu[2], Zhengyan Wan[3], ZheWang[4], Liyan Xie[5], DmitriyBespalov[6], Ying Nian Wu[1], Guang Cheng[1], Hongyuan Zha[7], 和 Qiang Sun[2,8]

[1]加利福尼亚大学洛杉矶分校[2]穆罕默德·本·扎耶德人工智能大学[3]华东师范大学[4]弗吉尼亚大学[5]明尼苏达大学[6]德雷塞尔大学[7]香港中文大学（深圳）[8]多伦多大学

摘要

掩码扩散模型已成为文本与多模态生成的强大框架。然而，其采样过程会同时更新多个词元，并将已生成的词元视为不可更改；当早期错误无法修订时，这可能导致误差累积。在本工作中，我们重新审视现有的自我纠正方法，并发现其局限源于额外的训练需求，或对失配似然估计的依赖。我们提出一种免训练自我纠正框架，利用预训练掩码扩散模型的归纳偏置。在不修改模型参数、也不引入辅助评估器的情况下，我们的方法以更少的采样步数，显著提升了文本到图像生成和多模态理解任务上的生成质量。此外，所提出的框架能够泛化至不同的掩码扩散架构，凸显了其稳健性与实际适用性。代码见 https://github.com/huge123/FreeCorrection。

1 引言

掩码扩散模型 [2, 13, 30, 22, 1, 26, 32, 29] 近来已成为文本 [24, 40, 5] 和多模态生成 [37, 38]的一种强大范式。通过掩码过程迭代地破坏数据，并学习逆转这一过程，掩码扩散模型能够实现并行生成，并为自回归解码提供了一种颇具吸引力的替代方案。越来越多的研究表明，这一框架能够随着模型规模和数据多样性的增长而良好扩展，并可无缝集成到大型预训练架构中。

尽管掩码扩散模型在实证中取得了成功，但它们仍存在一个尚未得到充分解决的根本局限。在每个反向步骤中，掩码扩散模型会同时更新多个词元；一旦生成了一个被掩码的词元，通常就会在后续步骤中将其视为不可更改。因此，在生成早期阶段引入的错误可能会持续存在并不断累积，导致合成分布与目标数据分布不匹配。从连续时间马尔可夫链（CTMC） [3]的角度，可以解释这一现象。掩码

扩散模型所采用的实际反向采样过程，在数学上等价于对底层反向时间连续时间马尔可夫链进行 τ-leaping 近似 [11]。这种近似将精确跳跃时刻边缘化，并对所有被掩码的维度执行并行的近似更新。尽管 τ-leaping 能够实现高效且高度并行的生成，但在连续时间马尔可夫链的数值分析中，众所周知，如果错误的跳跃过早发生且无法修正，就会引入路径级误差。

在本文中，我们重新审视了现有的旨在通过自我纠正来减轻生成错误的方法。以往工作通常依赖以下两种方式之一：（i）微调或重新训练额外模型，以显式评估或重新评分生成的词元 [32, 42, 15, 28, 14], ；或（ii）重复利用在反向过程的前几个步骤中获得的似然估计 [33]。尽管这些方法在某些情况下有效，但往往会带来额外的训练成本，并依赖不匹配的信号。相比之下，我们的目标是开发一种免训练的纠正机制，使其直接作用于反向过程，同时忠实于底层的生成动态。

我们的核心洞见是，掩码扩散网络本身已具备可用于自我纠正的有益归纳偏置，无需额外训练。通过仔细分析反向过程的结构及模型的条件预测，我们发现了更精细的设计选择，使模型能够在生成过程中重新评估并改进不确定的更新。重要的是，我们的方法无需修改预训练模型，也无需引入辅助评估器。

我们通过 GenEval 基准 [8]上的文本到图像生成和 VLMEvalKit [6] 上的多个多模态理解基准测试验证了我们的方法。据我们所知，这是首项研究多模态掩码扩散模型中自我纠正机制的工作。结果表明，所提方法显著提升了生成保真度和语义对齐能力，同时在不牺牲输出质量的情况下实现了更快的采样。此外，该方法能够泛化到不同的掩码扩散架构，凸显了其鲁棒性和实际适用性。结合对重新掩码准则设计的详细消融研究，我们的发现表明，利用预训练掩码扩散模型的归纳偏置进行重新掩码，为在推理时实现高效且有原则的自我纠正开辟了新的途径。

2 相关工作

掩码扩散模型。 掩码扩散模型（MDMs）通过采用掩码前向过程破坏数据，并学习相应的反向动力学，将去噪扩散范式推广到离散令牌空间。早期工作研究了针对离散变量的离散扩散，并为掩码破坏和并行解码建立了实用的训练与采样目标 [2, 13, 30, 22, 29]。近期研究进展主要聚焦于将掩码扩散扩展到大语言模型场景，并通过更优的调度方案和令牌选择规则（例如 Top-K updates [24, 40, 5]）提升解码效率/质量。

除了纯文本之外，MDMs 还通过首先采用 VQ-VAE [31] 将图像转换为离散令牌，然后通过大规模预训练 [37, 38]处理图像和文本令牌，扩展到多模态生成和统一视觉-语言建模。这些模型继承了扩散式并行生成的关键优势，但也加剧了离散并行采样中的一个已知问题：令牌一旦填入，通常就会被视为吸收态，导致早期错误难以修正。

掩码扩散模型的自我纠正。 越来越多的研究致力于赋予掩码扩散模型自我纠正能力，从而能够修改先前生成的词元。从连续时间马尔可夫链（连续时间马尔可夫链）的视角来看，一些方法采用预测器-校正器式采样，以减少反向动力学中的离散化误差 [3, 18, 7, 4, 27]。另一类

方法在反向过程中引入显式的重新掩码：词元可以根据调度方案重新进行掩码，然后再次生成，从而能够在采样过程中进行纠正 [33, 28]。

现有的自我纠正方法主要区别在于是否需要额外训练。一些方法对扩散过程进行泛化，并训练新模型以支持纠正和改进反向转移 [32, 28]。另一些方法通过辅助网络或微调来估计逐词元质量/置信度，然后根据这些估计决定要重新掩码或重新采样哪些词元 [42, 15, 14]。相比之下，我们的目标是利用预训练掩码扩散骨干网络中已有的归纳偏置，实现免训练 的多模态 掩码扩散模型自我纠正，同时不修改模型参数，也不引入外部评估器。

3 预备知识

3.1 掩码扩散模型的形式化定义

掩码扩散模型 [2, 13, 30, 22, 1, 26, 32, 29, 24] 由前向过程和反向过程构成。在前向过程中，我们逐步将句子中的每个词替换为一个特殊的掩码词元，直至所有词都被替换。在反向过程中，模型学习从全是掩码词元的状态开始，逐步补全句子。

在本文中，我们遵循[3, 7, 4, 20] 中使用的连续时间马尔可夫链（CTMC）框架来建模掩码扩散模型。我们考虑一个 $\mathcal{D}$维离散状态空间$\mathcal{S}^{\mathcal{D}}$，其中 $\mathcal{S}$ 表示每个维度的有限状态空间。数据分布记为 q_0，其在 $\mathcal{S}^{\mathcal{D}}$上的概率质量函数为 $q_0(x)$，数据样本记为 $x_0 \in \mathcal{S}^{\mathcal{D}}$。

前向过程。 对于给定的噪声水平 $t \in [0,1]$，前向过程$x_t \sim q_{t|0}\,(x_t|x_0)$是通过$q_{t|0}(x_t|x_0) = \prod_{d=1}^{\mathcal{D}} q_{t|0}^d(x_t^d|x_0^d)$实现的坐标独立掩码过程，其中

$$q_{t|0}^d(x_t^d|x_0^d) = (1-\alpha_t)\delta_m(x_t^d) + \alpha_t \delta_{x_0^d}(x_t^d),\ d \in [\mathcal{D}], \tag{1}$$

其中 α_t 是满足 $\alpha_0 = 1, \alpha_1 = 0$的预定义噪声调度； $\delta_x(z)$表示克罗内克δ函数，满足当 $x = z$ 时 $\delta_x(z) = 1$，当 $x \neq z$时 $\delta_x(z) = 0$； m 表示掩码的独热嵌入。换言之，对于每个坐标 i，在前向时间步 t， x_t^i 以概率 $1-\alpha_t$ 被替换为掩码词元，否则保持不变 [3]。

反向过程。上述前向过程在给定初始状态 x_0 条件下的预言机反向过程记为 $q_{s|t}\,(x_s|x_t,\ x_0) = \prod_{d=0}^{D} q_{s|t}\left(x_s^d|x_t,\ x_0\right)$，对任意 $s < t$，其中

$$q_{s|t}\left(x_s^d|x_t, x_0\right) = \begin{cases} \delta_{x_t^d}(x_s^d), & \text{if } x_t^d \neq m, \\ \frac{1-\alpha_s}{1-\alpha_t}\delta_m(x_s^d) + \frac{\alpha_s-\alpha_t}{1-\alpha_t}\delta_{x_0^d}(x_s^d), & \text{if } x_t^d = m. \end{cases} \tag{2}$$

训练目标。 一个由 θ 参数化的神经网络 p_θ，可以通过最小化真实反向条件分布 $q_{s|t}\left(x_s^d|x_t, x_0\right)$ 与 $q_{s|t}\left(x_s^d|x_t, p_\theta^d(x_t,t)\right)$之间的 Kullback–Leibler（KL）散度，来训练其近似预言机反向过程。其中， $p_\theta^d(x_t,t)$ 表示网络输出的第 d 个坐标。训练目标可概括为

$$\mathcal{L}_\theta = \int_0^1 \frac{\alpha_t'}{1-\alpha_t} \mathop{\mathbb{E}}_{\substack{x_0\sim q_0 \\ x_t\sim q_{t|0}(x_t|x_0)}} \left[\sum_{d=1}^{D} \delta_m(x_t^d) {x_0^d}^\top \log p_\theta^d\,(x_t,t)\right] dt.$$

这里， $\alpha'_t \triangleq \frac{d\alpha_t}{dt}$。在实践中，通常采用无时间嵌入架构的去噪网络，即 $p_\theta(x_t, t) = p_\theta(x_t)$，因为 x_t 会通过掩码词元的数量隐式包含关于 t 的信息。

采样过程。 在学习网络 $p_\theta(x_t)$ 后，可以通过模拟公式 (2) 中的反向过程来生成样本。由于反向动力学在各维度之间可分解，因此可以针对每个坐标独立采样。具体而言，对于时间 t 处的每个掩码位置 d，在区间 $(s, t]$ 内，会以概率 $\frac{\alpha_s - \alpha_t}{1-\alpha_t}$ 从掩码词元发生转移；在发生此类转移的条件下，新词元从分类分布 $p^d_\theta(x_t)$ 中采样。

3.2 掩码扩散模型的局限性

为了更好地理解掩码扩散模型采样过程的局限性，下面我们将详细讨论坐标独立的非齐次连续时间马尔可夫链。

首先可以观察到，对于每个维度 d，前向掩码过程具有以下时变转移率：

$$Q_t^{(f)}(x_0^d \to m) = \lambda(t) := -\frac{\alpha'_t}{\alpha_t},$$

因为由此得到的转移概率满足

$$q(x_t^d = x_0^d | x_0^d) = \exp\left(-\int_0^t \lambda(\tau)\, d\tau\right) = \alpha_t,$$

这恢复了公式 (1) 中预定义的噪声调度方案。

反向过程对应于时间反向连续时间马尔可夫链，它会逐步将掩码状态转换回来自数据分布的采样。反向时间转移率可以通过两种等价方式计算。一种方法（ [3, 23], 采用的方法）将反向速率矩阵 $Q_t^{(r)}$ 表示为正向速率的形式：

$$Q_t^{(r)}(x_t \to x_s) = Q_t^{(f)}(x_s \to x_t) \sum_{x_0} \frac{q_{t|0}(x_s|x_0)}{q_{t|0}(x_t|x_0)}\, q_{0|t}(x_0|x_t), x_t \neq x_s. \tag{3}$$

或者，按照 [7, 4], 的方法，可以通过对数据状态后验分布进行边缘化来得到反向转移率：

$$Q_t^{(r)}(x_t \to x_s) = \mathbb{E}_{q_{0|t}(x_0|x_t)}\left[Q_t^{(r)}(x_t \to x_s|x_0)\right],$$

其中，数据状态后验分布为 $q_{0|t}(x_0 \mid x_t) = \frac{q_{t|0}(x_t|x_0)\ q_0(x_0)}{\mathbb{E}_{q_0(x_0)}[q_{t|0}(x_t|x_0)]}$，而 $Q_t^{(r)}(x_t \to x_s|x_0)$ 表示给定初始状态 x_0 时反向时间连续时间马尔可夫链的条件转移率。计算后，我们得到反向转移率：

$$Q_t^{(r)}(x_t \to x_s) = \frac{\alpha'_t}{1-\alpha_t}\delta_m(x_t) p_{0|t}\left(x_0 = x_s|x_t\right).$$

反向连续时间马尔可夫链可以使用 Gillespie算法 [9, 10, 36], 进行精确模拟：该算法从指数分布中采样下一次跳跃时间，并选择单一维度进行更新。正如 [3], 所指出的，对于较大的 D，这种方法效率低下。在实践中，各种方法 [3, 23]采用 τ-Leaping [11, 36] 作为并行算法，在较短的时间间隔内同时更新所有转移

。第3.1节中描述的采样过程，是对连续时间马尔可夫链采样中使用τ-Leaping的一阶近似，如下所示。

有限区间 $[s,t]$ 上的精确转移概率可通过对转移率进行积分得到。由于 CTMC 在解除掩码后即处于吸收态，因此只有从 m 到数据词元 v 的转移是非平凡的。记 $x_t^d = m$，对于任意 $\tau \in [s,t]$，该转移概率满足以下柯尔莫哥洛夫前向方程：

$$\frac{d}{d\tau} q_{\tau|s}(v|m) = Q_\tau^{(r)}(m \to v) = \frac{\alpha'_\tau}{1-\alpha_\tau}\, p_{0|\tau}(v|m).$$

因此，转移概率可以写为

$$\begin{aligned} q_{s|t}(x_s^d = v \mid x_t^d = m) &= \int_s^t \frac{\alpha'_\tau}{1-\alpha_\tau}\, p_{0|\tau}(v \mid m)\, d\tau \\ &\approx p_{0|t}(v \mid x_t, t) \int_s^t \frac{\alpha'_\tau}{1-\alpha_\tau}\, d\tau \\ &= p_{0|t}(v \mid x_t, t)\, \log \frac{1-\alpha_t}{1-\alpha_s} \\ &\approx p_{0|t}(v \mid x_t, t) \frac{\alpha_s - \alpha_t}{1-\alpha_t}. \end{aligned}$$

因此，我们证明了掩码扩散模型的采样过程是底层反向时间连续时间马尔可夫链的 τ-leaping 采样的一阶近似。

τ-leaping 的一个典型特性是，同时执行跨越多个维度的近似跃迁。由于掩码扩散模型的特殊设计，一旦某个坐标从掩码词元转移到数据词元，该坐标便进入吸收态。因此，早期并行跃迁引入的误差无法在后续步骤中得到纠正。为了缓解并行生成导致的误差累积，研究者提出了两条主要思路。

第一条思路旨在将生成限制在模型具有较高置信度的词元上。具体而言，Top-K策略 [43, 39, 34, 25, 35, 37]已被现代掩码扩散模型广泛采用。Top-K解码不再将每个位置视为以概率 ($\alpha_s - \alpha_t$)/(1 − α_t)生成，而是根据最高的预测令牌概率选择要更新的位置。Top-K边际策略 [16]采用前两个预测概率之差。然而，这类方法只能缓解并行生成所造成的误差。

另一类工作赋予掩码扩散模型自我纠正能力 [3, 18, 7,4, 33, 28, 32, 42, 15, 27], ，使先前生成的词元能够在后续步骤中被修订。

3.3 掩码扩散模型的自我纠正

[3, 18, 7, 4] 针对连续时间马尔可夫链采用预测校正法。 [33, 28] 推广反向转移概率，使已生成的词元可按调度方案 σ_t被重新掩码，而 [32] 则为实现自我纠正同时推广正向过程与反向过程，这需要训练一个新的掩码扩散模型。 [42, 15, 14] 将逐词元质量定义为重新掩码的准则，并微调预训练模型以估计逐词元质量。我们在表1中根据现有自我纠正方法是否需要微调现有模型或训练新模型进行了总结。

4 所提方法

本文旨在探究预训练多模态掩码扩散模型能否在无需额外微调或重新训练的情况下实现自我纠正。为了实现免训练

表1：掩码扩散模型自我纠正的分类。

类别	参考文献
免训练	[33]
基于训练或基于微调	[32, 42, 15, 28, 14]

表2：GenEval 基准上的文本到图像生成评估。

方法	单一	Two	计数	颜色	Pos.	属性	总体 ↑
Lumina-DiMOO [a]	0.99	0.93	0.85	0.84	0.84	0.71	0.86
Lumina-DiMOO（ReMDM）	**1.00**	**0.94**	0.86	0.87	0.82	0.74	0.87
Lumina-DiMOO（我们的方法）	0.99	**0.94**	**0.88**	**0.93**	**0.87**	**0.79**	**0.90**

[a]为进行公平比较，我们在 GenEval 基准上重新评估了预训练的 Lumina-DiMOO。生成步骤数为 64。

自我纠正，我们需要重新审视 [33] 提出的现有免训练框架。他们将式（2）中的掩码扩散模型反向过程推广为

$$q_\sigma\left(x_s|x_t,x\right)=\begin{cases}(1-\sigma_t)\delta_{x_t}(x_s)+\sigma_t\delta_m(x_s), & x_t\neq m,\\ \frac{\alpha_s-(1-\sigma_t)\alpha_t}{1-\alpha_t}\delta_x(x_s)+\frac{1-\alpha_s-\sigma_t\alpha_t}{1-\alpha_t}\delta_m(x_s), & x_t=m.\end{cases}\tag{4}$$

也就是说，词元生成后，仍有 σ_t 的概率会被重新掩码。[33] 提出了三种重新掩码调度方案，分别为 σ_t，并发现采用重缩放的基于置信度的调度方案效果最佳。在该调度方案中，生成时的预测令牌概率被用作重新掩码的依据。预测概率较低的词元更有可能被重新掩码。从合理性上看，这一依据是成立的，因为预测令牌概率 $p_\theta^d\left(x_t\right)$ 仅在 $x_t^d=m$ 时有效。

然而，这一依据存在局限：我们需要根据当前 x_t （而非前一时间步，即词元生成时）评估每个位置的置信度。因此，近期研究 [42, 15] 提出在每个解码步骤重新评估每个已生成位置的预测令牌概率，即 $p\left(x_t^d\right)=p_\theta^d\left(x_t^d|x_t^{\backslash d}\right)$，其中 $x_t^{\backslash d}$ 表示将 x_t 的第 d 维改为 m，而 $p_\theta^d\left(x_t^d|\cdot\right)$ 表示掩码扩散模型对词元 x_t^d 的预测概率。[42] 提出训练一个空心 Transformer，以实现高效估计；[15] 则提出微调预训练掩码扩散模型来估计预测令牌概率。由此自然引出一个问题：鉴于预训练掩码扩散模型的归纳偏置，我们能否免训练地估计预测令牌概率？

4.1 免训练自我纠正的合理性

在本小节中，我们使用真实实验结果来验证预训练掩码扩散模型的归纳偏置。我们采用一种先进的多模态掩码扩散模型 [37]。我们旨在探究预训练网络在生成过程中是否能够识别错误词元。具体而言，我们从 Flickr8k 数据集 [12] 中采样多个图像-文本提示对（z_0^{image}，z^{prompt}）。使用预训练的语言和图像分词器，我们将（z_0^{image}，z^{prompt}）投影到嵌入空间（x_0，y）中。接着，我们应用掩码扩散模型的加噪过程 $q_{t|0}(x_t|x_0)$，以获得带噪表示 x_t，其中 $t\sim U(0,1)$ 和 $U(0,1)$ 表示单位区间 $(0,1)$ 上的均匀分布。随后，我们随机选取未掩码

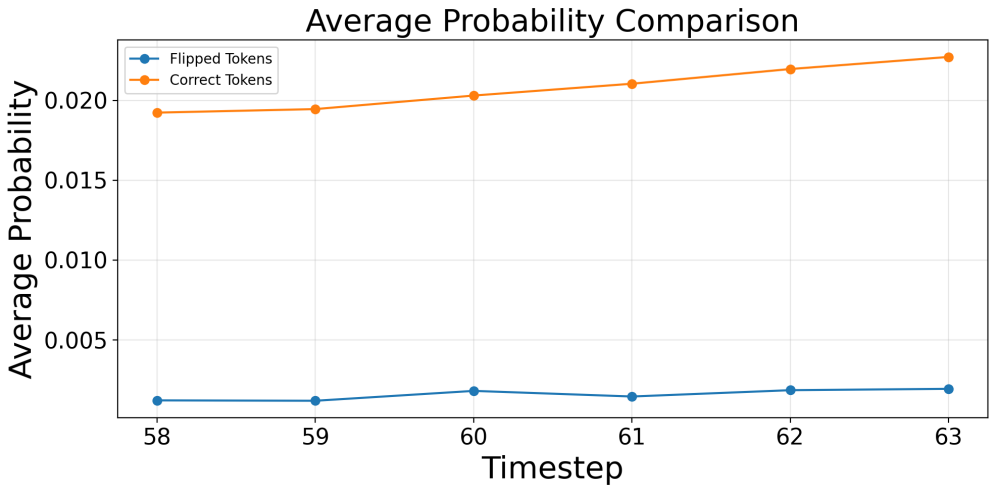

图 1：在 2000 个采样中，翻转词元和正确词元的平均预测概率。横轴表示生成的时间步（文本到图像生成共 64 步），纵轴表示所有翻转位置和正确位置上的平均概率。

词元并人为翻转其值。这些操纵词元被输入预训练网络，我们通过检查模型赋予它们的似然，评估模型是否能够检测出这些翻转词元。

我们在图 1 中绘制了原始词元及其翻转对应词元在不同时间步的平均似然值。结果表明，预训练掩码扩散模型会为更合理的词元（校正词元）分配显著高于不合理词元（翻转词元）的预测概率。这一现象反映了预训练模型对已生成词元的归纳偏置。

在图 2 中，我们展示了累积预测概率能够很好地区分错误词元。对于时间步 32 之后的每个时间步，我们发现，错误词元的累积概率平均秩与单一时间步评估所得的平均秩相比，差距越来越大。因此，这启发我们设计一种用于自我纠正的滑动窗口。

4.2 自校正准则的深入设计

考虑到预训练掩码扩散模型的归纳偏置，我们提出使用预测概率 $p_\theta^d(x_t)$进行重新掩码，即使 x_t 已经生成。该设计使模型能够在整个生成过程中充分发挥其能力。

基于这一表述，重新掩码策略存在多种设计选择。我们在下文总结几个关键方面：

- **重掩码规则**：重新掩码既可以通过应用硬规则（例如，类似于生成过程中所使用的Top-K 策略）来执行，也可以通过对由预测概率 $p_\theta^d(x_t)$定义的多项分布进行采样来执行。

- **时序聚合：**重新掩码决策可能取决于跨生成步骤对预测概率 $p_\theta^d(x_t)$的累积。

- **Top-K 重新掩码边际：**生成中使用的Top-K边际策略是否也能为重新掩码带来改进

算法 1 掩码扩散模型的免训练自我纠正

要求： 预训练去噪网络 $\mathbf{x}_\theta$，步数 T，噪声调度 $\{\alpha_t\}$，重新掩码调度方案 $\{\sigma_t\}$，重新掩码评分类型 Score $\in$ {当前，累积}，重新掩码规则 Rule $\in$ {确定性，随机性}

确保： 采样$\mathbf{x}_0$

初始化 $\mathbf{x}_T \leftarrow \boldsymbol{m}$

初始化逐词元评分 $\mathbf{S} \leftarrow \mathbf{0}$

对于 $i = T, T-1, \ldots, 1$ **do**

$t \leftarrow i/T, \quad s \leftarrow (i-1)/T$

检索 α_t，α_s，σ_t

（1）标准反向更新

$\hat{\mathbf{x}} \leftarrow \mathbf{x}_\theta(\mathbf{x}_t)$

在式（4）中采样 $\mathbf{x}_s \sim q_\sigma(\mathbf{x}_s \mid \mathbf{x}_t, \hat{\mathbf{x}})$ {}

$\mathbf{x}_t \leftarrow \mathbf{x}_s$

（2）免训练似然评估

对每个位置 $d \in [D]$ with $x_t^d \neq m$ **执行**

$\ell_t^d \leftarrow p_\theta^d\left(x_t^d \mid \mathbf{x}_t\right)$ {重新评估似然}

如果 分数 = 当前 **则**

$S^d \leftarrow \ell_t^d$

else

$S^d \leftarrow S^d + \log(\ell_t^d)$ {累积似然}

结束 if

结束 for

(3) 重新掩码步骤

$K_t \leftarrow \lfloor \sigma_t \cdot D \rfloor$

如果规则 = 确定性**那么**

$\mathcal{R} \leftarrow \arg\min_{|\mathcal{R}|=K_t} \{S^d\}_{d:x_t^d \neq m}$ {重新掩码 K_t 得分最低的令牌}

else

采样 $\mathcal{R}$ 大小为 K_t，概率为 $\propto \exp(-S^d)$，从 $\{d$ 中采样：$x_t^d \neq m\}$

结束如果

对于每个 $d \in \mathcal{R}$ **执行**

$x_t^d \leftarrow m$

结束循环

结束循环

返回 $\mathbf{x}_0$

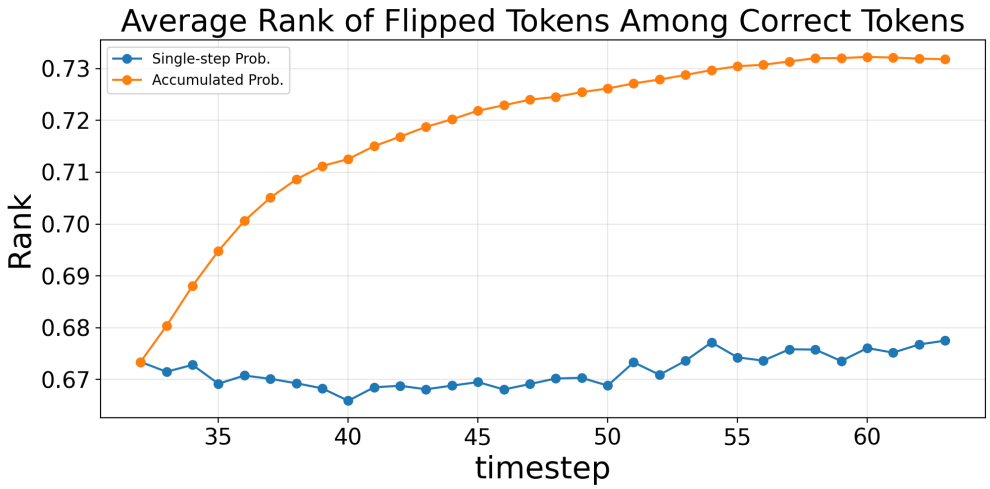

图 2：使用累积预测概率的有效性。横轴表示生成的时间步，纵轴表示翻转词元的预测概率在正确词元中的平均秩。秩越大，概率越小。

- **基于分布不确定性的重新掩码**：受近期工作 [17], 启发，我们考虑利用预测概率的完整词汇分布中的更丰富信息$p_\theta^d(x_t)$，而不是仅依赖 Top-K 或 Top-K 边际。具体而言，我们使用连续时间步的预测分布之间的 KL 散度或 Wasserstein 距离作为衡量模型在每个位置上置信度的标准。生成词元表现出更大分布变化的位置被视为置信度较低。因此，在采样过程中，我们对距离最大的 Top-K 个位置重新掩码。

我们在算法 1 中总结了总体流程。在下一节中，我们将仔细展示所提出框架的优越性，并详细比较重新掩码策略的各种选择。

5 实验

实验设置 我们使用最先进的多模态掩码扩散模型 [37] 开展实验，并且仅在推理时应用我们的自我纠正策略，无需进行任何额外的微调或重新训练。为评估所提出的重新掩码策略的有效性，我们在文本到图像生成和多模态理解任务上测试其表现。对于文本到图像生成，我们采用广泛使用的 GenEval 基准 [8], ，该基准侧重于涉及物体计数、空间关系和属性绑定的复杂组合提示。对于多模态理解，我们报告 MMBench（MMB） [21], 、SEED-Bench [19], 和 MMMU [41]上的结果。对于所有多模态理解基准测试，我们将每个问题的最大生成长度设为 256 个词元。为获得自动且一致的正确性判断，我们采用 GPT-4o-mini 作为验证器：给定模型回答和基准提供的参考答案，验证器会输出判断结果。除非另有说明，我们使用与原始模型相同的骨干网络、分词器和评估流程，仅通过插入所提出的重新掩码步骤来改变采样过程。



图3：Lumina-DiMOO、ReMDM 和我们的方法生成图像的比较。我们的方法以更少的采样步数实现了更高的质量。

表 3：在多个基准测试上评估多模态理解。

模型	MMB ↑	SEED ↑	MMMU ↑
Lumina-DiMOO [a]	58.7	71.4	41.4
Lumina-DiMOO (ReMDM)	57.8	**74.3**	43.4
Lumina-DiMOO（我们的方法）	**60.7**	74.2	**44.0**

[a]为进行公平比较，我们在多个多模态理解基准测试上重新评估了预训练的 Lumina-DiMOO。生成步数为 64，生成序列长度为 256。

5.1 总体有效性

我们首先评估所提出的重新掩码策略的总体效果。如表 2 所示，与原始 Lumina-DiMOO 以及现有的免训练自我纠正方法相比，我们的方法提升了 GenEval 表现 [33]。在表 3 和图 5 所示的多模态理解任务中，我们观察到，在推理时启用基于重新掩码的自我纠正后，性能均有提升。总体而言，这些结果支持了我们的核心主张：我们可以利用预训练掩码扩散模型的归纳偏置，识别并修正低置信度词元，从而在无需额外训练的情况下提升最终生成质量。

5.2 重新掩码设计选择的消融研究

我们进行了全面的消融研究，以分析第 4.2 节中所述不同重新掩码设计选择的影响。具体而言，我们在表 4 中比较了各种重新掩码规则、时间聚合策略和不确定性标准，以更好地了解它们各自的作用。总体而言，“累积似然 + 确定性重新掩码”表现最佳。这证实了累积预测概率有助于识别错误的

表 4：GenEval 基准上不同重新掩码策略设计的消融研究。“确定性重新掩码”指采用贪心算法，对预测概率最小的 K 个位置重新掩码；“随机重新掩码”则指根据预测概率对需要重新掩码的位置进行采样。预测概率越小，该位置越有可能被重新掩码。

方法	单一	Two	计数	颜色	Pos.	属性	总体 ↑
Lumina-DiMOO	0.99	0.93	0.85	0.84	0.84	0.71	0.86
+当前步似然 +确定性重新掩码	0.95	0.92	0.83	0.87	0.84	0.76	0.86
+当前步似然 +随机重新掩码	0.96	0.94	0.86	0.86	0.84	0.75	0.87
+ 累积似然 +确定性重新掩码	0.99	0.94	0.88	**0.93**	**0.87**	**0.79**	**0.90**
+ 累积似然 +随机重新掩码	0.99	0.93	0.85	0.88	0.82	0.78	0.87
+Top K 间隔	0.99	0.93	**0.90**	0.89	0.86	0.75	0.89
+KL 散度	0.98	**0.95**	0.85	0.86	0.82	0.72	0.86
+ Wasserstein 距离	**1.00**	0.93	0.84	0.88	0.84	0.66	0.86

表 5：基于 MMaDA-8B-MixCoT [38] 在 GenEval 基准上对所提方法进行评估。

方法	单一	Two	计数	颜色	Pos.	属性	总体 ↑
MMaDA-8B-MixCoT [1] (Top K)	**0.93**	0.47	0.31	0.81	0.16	0.26	0.49
MMaDA-8B-MixCoT（Top K 间隔）	0.88	0.49	0.29	0.79	0.15	**0.28**	0.48
MMaDA-8B-MixCoT（ReMDM）	0.91	0.59	0.38	0.78	**0.20**	0.19	0.51
MMaDA-8B-MixCoT（我们的方法）	0.91	**0.67**	**0.39**	**0.84**	0.15	0.18	**0.52**

为进行公平比较，我们在 GenEval 基准上重新评估了预训练的 MMaDA-8B-MixCoT。生成步骤为 64。

词元，如图 2 所示。这也验证了广泛使用的 Top-K 策略的有效性。我们针对表 4 所示的性能进行了细致的超参数搜索。超参数“重新掩码调度方案” σ_t 与要重新掩码的词元数量 K_t 相关，其中 $K_t = \lfloor \sigma_t \cdot D \rfloor$，而 $\lfloor \cdot \rfloor$ 表示向下取整函数。我们根据验证集上的表现确定了该超参数。

5.3 高效采样的重新掩码

接下来，我们研究所提出的重新掩码机制能否在推理时作为一种实用的加速技术。在保持其他设置不变的情况下，我们调整反向步骤的数量，并报告 GenEval 总体得分。如图 4 所示，仅使用 16 个采样步数时，我们的方法取得了与 64 步 Lumina-DiMOO 基线相当或更好的 GenEval 表现。这些结果表明，有选择地重新处理低置信度位置可以弥补早期步骤中的错误，并在不牺牲生成质量的前提下实现四倍加速。与 ReMDM 相比，所提方法带来的提升更大，这证明了利用预训练模型归纳偏置的有效性。我们还在图 3 中展示了一些生成图像的可视化结果。

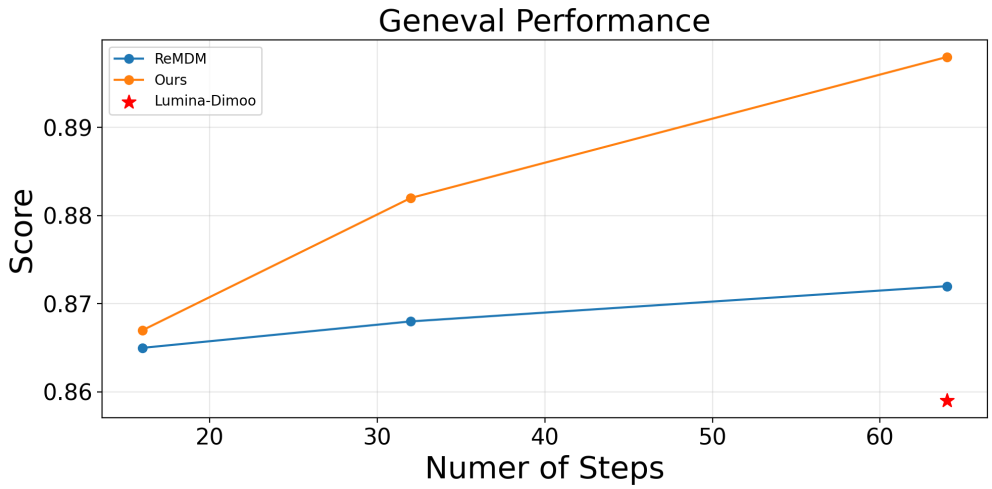

图 4：不同推理步数下的 GenEval 表现。横轴表示生成步数，纵轴表示 GenEval 基准上的总体表现。



Which one is the correct caption of this image?

A: People in a horse drawn buggy on a city street.

B: A fire hydrant with a pair of eye stickers making a face on it.

C: a large food truck is parked on the side of the street

D: Neither one of these people had a good flight.

Lumina-DiMOO: Long response repeat the question several times.

ReMDM: Long response repeat the question several times.

Ours: To determine the correct caption for the image,
we need to carefully observe the details present in the image.
The image shows a fire hydrant with a pair of eye stickers on it.

图 5：多模态理解表现示意图。

5.4 跨骨干模型的泛化能力

最后，我们评估重新掩码策略是否依赖于特定骨干网络，还是能够推广到不同的掩码扩散架构。在表 5 中，我们评估了所提方法在 MMaDA-8B-MixCoT [38] 上的文本到图像生成性能。我们发现，启用相同的推理时自校正规则后，性能均能稳定提升。这表明我们的方法主要利用了掩码扩散模型的通用归纳偏置，凸显了该方法的稳健性。

6 结论

在本文中，我们解释了并行且不可逆的词元更新所导致的误差累积现象。随后，我们提出了一种模型无关的免训练自我纠正框架，可直接在推理时运行。与依赖额外训练、辅助评估器或失准似然的先前方法不同，我们的方法利用预训练掩码扩散网络的归纳偏置进行自我纠正。在文本到图像生成和多模态理解基准测试上进行的大量实验表明，我们的方法能够持续提升生成质量，同时加快采样速度而不牺牲输出质量。观察到的提升可泛化至不同的掩码扩散架构，彰显了所提方法的稳健性和实际适用性。

参考文献

[1] Marianne Arriola、Aaron Gokaslan、Justin T Chiu、Zhihan Yang、Zhixuan Qi、Jiaqi Han、Subham Sekhar Sahoo和Volodymyr Kuleshov。块扩散：在自回归与扩散语言模型之间进行插值。*ICLR*，2025。[2] Jacob Austin、Daniel D. Johnson、Jonathan Ho、Daniel Tarlow和Rianne van den Berg。离散状态空间中的结构化去噪扩散模型。见*NeurIPS*，2021。[3] Andrew Campbell、Joe Benton、Valentin De Bortoli、Thomas Rainforth、George Deligiannidis和Arnaud Doucet。离散去噪模型的连续时间框架。见神经信息处理系统进展大会，第35卷，第28266–28279页，2022。[4] Andrew Campbell、Jason Yim、Regina Barzilay、Tom Rainforth和Tommi Jaakkola。离散状态空间上的生成流：实现多模态流及其在蛋白质协同设计中的应用。见国际机器学习会议，2024。[5] Shuang Cheng、Yihan Bian、Dawei Liu、Linfeng Zhang、Qian Yao、Zhongbo Tian、Wenhai Wang、Qipeng Guo、Kai Chen、Biqing Qi和Bowen Zhou。Sdar：一种用于可扩展序列生成的协同扩散-自回归范式。*ArXiv*，abs/2510.06303，2025。[6] Haodong Duan、Junming Yang、Yu Qiao、Xinyu Fang、Lin Chen、Yuan Liu、Xiao wen Dong、Yuhang Zang、Pan Zhang、Jiaqi Wang、Dahua Lin和Kai Chen。Vlmevalkit：用于评估大型多模态模型的开源工具包。第*32*届*ACM*国际多媒体会议论文集，2024。[7] Itai Gat、Tal Remez、Neta Shaul、Felix Kreuk、Ricky T. Q. Chen、Gabriel Synnaeve、Yossi Adi和Yaron Lipman。离散流匹配。*NeurIPS*，2024。[8] Dhruba Ghosh、Hanna Hajishirzi和Ludwig Schmidt。Geneval：一种以对象为中心的文图对齐评估框架。*NeurIPS*数据集与基准测试专题，2023。[9] Daniel T. Gillespie。一种数值模拟耦合化学反应随机时间演化的通用方法。计算物理学杂志，22:403–434，1976。[10] Daniel T. Gillespie。耦合化学反应的精确随机模拟。物理化学杂志，81:2340–2361，1977。[11] Daniel T. Gillespie。化学反应体系的近似加速随机模拟。化学物理学杂志，115:1716–1733，2001。

[12] Micah Hodosh、Peter Young和Julia Hockenmaier。将图像描述构建为排序任务：数据、模型与评估指标。人工智能研究期刊，47:853– 899，2013。[13] Emiel Hoogeboom、Didrik Nielsen、Priyank Jaini、Patrick Forr'e和Max Welling。Argmax流与多项式扩散：迈向非自回归语言模型。*NeurIPS*，2021。[14] Zemin Huang、Yuhang Wang、Zhiyang Chen和Guo-Jun Qi。不要过早定论：面向扩散语言模型的自反思重新掩码。*ArXiv*，abs/2509.23653，2025。[15] Jaeyeon Kim、Seunggeun Kim、Taekyun Lee、David Z. Pan、Hyeji Kim、Sham M. Kakade和Sitan Chen。面向可证明自我纠正的掩码扩散微调。*ArXiv*，abs/2510.01384，2025。[16] Jaeyeon Kim、Kulin Shah、Vasilis Kontonis、Sham M. Kakade和Sitan Chen。面向最坏情况训练，面向最佳情况规划：理解掩码扩散中的词元排序。*ICML*，2025。[17] Seo Hyun Kim、Sunwoo Hong、Hojung Jung、Young Ju Park和Se-Young Yun。Klass：掩码扩散模型中KL引导的快速推理。*NeurIPS*，2025。[18] Jos´e Lezama、Tim Salimans、Lu Jiang、Huiwen Chang、Jonathan Ho和Irfan Essa。用于图像合成的离散预测-校正扩散模型。见学习表征国际会议，2023。[19] Bohao Li、Rui Wang、Guangzhi Wang、Yuying Ge、Yixiao Ge和Ying Shan。Seed-bench：基于生成式理解对多模态大语言模型进行基准测试。*CVPR*，2024。[20] Sulin Liu、Juno Nam、Andrew Campbell、Hannes St¨ark、Yilun Xu、T. Jaakkola和Rafael G'omez-Bombarelli。边生成边思考：带规划去噪的离散扩散。*ICLR*，2025。[21] Yuanzhan Liu、Haodong Duan、Yuanhan Zhang、Bo Li、Songyang Zhang、Wangbo Zhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu、Kai Chen和Dahua Lin。Mmbench：你的多模态模型是全能型选手吗？*ArXiv*，abs/2307.06281，2023。[22] Aaron Lou、Chenlin Meng和Stefano Ermon。通过估计数据分布的比率进行离散扩散建模。见国际机器学习会议，2023。[23] Aaron Lou、Chenlin Meng和Stefano Ermon。通过估计数据分布的比率进行离散扩散建模。见国际机器学习会议，2024。[24] Shen Nie、Fengqi Zhu、Zebin You、Xiaolu Zhang、Jingyang Ou、Jun Hu、Jun Zhou、Yankai Lin、Ji-Rong Wen和Chongxuan Li。大语言扩散模型。*NeurIPS*，2025。[25] Shen Nie、Fengqi Zhu、Zebin You、Xiaolu Zhang、Jingyang Ou、Jun Hu、Jun Zhou、Yankai Lin、Ji-Rong Wen和Chongxuan Li。大语言扩散模型。见学习表征国际会议，2025。[26] Jingyang Ou、Shen Nie、Kaiwen Xue、Fengqi Zhu、Jiacheng Sun、Zhenguo Li和Chongxuan Li。你的吸收离散扩散在隐式地建模干净数据的条件分布。*ICLR*，2025。[27] Yonghyun Park、Chieh-Hsin Lai、Satoshi Hayakawa、Yuhta Takida和Yuki Mitsufuji。跳跃步长：优化离散扩散模型的采样调度。*ICLR*，2025。

[28] Zhangzhi Peng、Zachary Bezemek、Sawan Patel、Jarrid Rector-Brooks、Sherwood Yao、Alexander Tong 和 Pranam Chatterjee。掩码扩散模型采样的路径规划。*ArXiv*, abs/2502.03540，2025。[29] Jiaxin Shi、Kehang Han、Zhe Wang、Arnaud Doucet 和 Michalis K. Titsias。针对离散数据的简化广义掩码扩散。*NeurIPS*，2024。[30] Haoran Sun、Lijun Yu、Bo Dai、Dale Schuurmans 和 Hanjun Dai。基于分数的连续时间离散扩散模型。*ICLR*，2023。[31] A¨aron van den Oord、Oriol Vinyals 和 Koray Kavukcuoglu。神经离散表示学习。载于神经信息处理系统，2017。[32] Dimitri von Rutte、Janis Fluri、Yuhui Ding、Antonio Orvieto、Bernhard Scholkopf 和 Thomas Hofmann。广义插值离散扩散。*ICML*, 2025。[33] Guanghan Wang、Yair Schiff、Subham Sekhar Sahoo 和 Volodymyr Kuleshov。通过推理时扩展重新掩码离散扩散模型。*NeurIPS*，2025。[34] Xinyou Wang、Zaixiang Zheng、Fei Ye、Dongyu Xue、Shujian Huang 和 Quanquan Gu。扩散语言模型是多才多艺的蛋白质学习器。*ICML*，2024。[35] Yinjie Wang、Ling Yang、Bowen Li、Ye Tian、Ke Shen 和 Mengdi Wang。革新扩散大语言模型的强化学习框架。2026。[36] Darren J. Wilkinson。系统生物学的随机建模。2006。[37] Yi Xin、Qi Qin、Siqi Luo、Kaiwen Zhu、Juncheng Yan、Yan Tai、Jiayi Lei、Yuewen Cao、Ke Wang、Yibin Wang、Jinbin Bai、Qian Yu、Dengyang Jiang、Yuandong Pu、Haoxing Chen、Le Zhuo、Junjun He、Gen Luo、Tian-Xin Li、Ming Hu、Jin Ye、Shenglong Ye、Bo Zhang、Chang Xu、Wenhai Wang、Hongsheng Li、Guangtao Zhai、Tianfan Xue、Bin Fu、Xiaohong Liu、Yu Qiao 和 Yihao Liu。Lumina-dimoo：用于多模态生成与理解的全能扩散大语言模型。*ArXiv*, abs/2510.06308，2025。[38] Ling Yang、Ye Tian、Bowen Li、Xinchen Zhang、Ke Shen、Yunhai Tong 和 Mengdi Wang。MMaDA：多模态大型扩散语言模型。*arXiv*预印本 *arXiv:2505.15809*，2025。[39] Jiacheng Ye、Jiahui Gao、Shansan Gong、Lin Zheng、Xin Jiang、Zhenguo Li 和 Lingpeng Kong。超越自回归：用于复杂推理与规划的离散扩散。*ICLR*, 2025。[40] Jiacheng Ye、Zhihui Xie、Lin Zheng、Jiahui Gao、Zirui Wu、Xin Jiang、Zhenguo Li 和 Lingpeng Kong。Dream 7b：扩散大语言模型。*ArXiv*，abs/2508.15487, 2025。[41] Xiang Yue、Yuansheng Ni、Kai Zhang、Tianyu Zheng、Ruoqi Liu、Ge Zhang、Samuel Stevens、Dongfu Jiang、Weiming Ren、Yuxuan Sun、Cong Wei、Botao Yu、Ruibin Yuan、Renliang Sun、Ming Yin、Boyuan Zheng、Zhenzhu Yang、Yibo Liu、Wenhao Huang、Huan Sun、Yu Su 和 Wenhu Chen。MMMU：面向专家级通用人工智能的海量多学科多模态理解与推理基准。*2024 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第9556–9567页，2023。[42] Yixiu Zhao、Jiaxin Shi、Lester Mackey 和 Scott W. Linderman。离散扩散模型的知情校正器。*NeurIPS*，2025。

[43] Lin Zheng、Jianbo Yuan、Lei Yu 和 Lingpeng Kong。一种用于文本生成的重参数化离散扩散模型。*COLM*，2024。