arXiv:2602.01362v1 [cs.CL] 1 Feb 2026

平衡 g 离散扩散模型中的理解与生成 s

Yue Liu [1] Yuzhong Zha o [1] Zheyong Xie [2] Qixiang Ye [1] Jianbin Jiao [1] Yao Hu [2] Shaosheng Cao [2] Yunfan Liu [1]

摘要

在离散生成式建模中，两种主流范式展现出截然不同的能力：掩码扩散语言模型（MDLM）擅长语义理解和零样本泛化，而均匀噪声扩散语言模型（UDLM）能够实现出色的少步生成质量，但两者都无法在这两个维度上取得均衡表现。为此，我们提出XDLM，通过平稳噪声核连接这两种范式。XDLM 有两项关键贡献：（1）为 MDLM 和 UDLM 提供了有理论依据的统一框架，并将两种范式分别还原为特例；（2）通过对后验概率进行代数化简，缓解了内存瓶颈。实验表明，XDLM 推进了理解能力与生成质量之间的帕累托前沿。在量化结果上，XDLM 在零样本文本基准测试中比 UDLM 高出 5.4 分，并在少步图像生成方面优于 MDLM（FID 为 54.1，而 MDLM 为 80.8 ）。将 XDLM 扩展用于微调一个 8B 参数的大语言模型后，仅用 32 步便达到 15.0 MBPP，基线性能几乎翻倍。最后，对训练动态的分析揭示了 XDLM 在长期扩展方面的卓越潜力。代码见此 URL。

1. 引言

扩散模型在连续域中取得了显著成功，尤其是在图像和音频生成方面（Ho 等人，2020年；Dhariwal & Nichol，2021年；Rombach 等人，2022年；Kong 等人，2020年）。受这一潜力启发，离散去噪扩散概率模型（D3PM）已成为向离散状态空间转变的一种颇具前景的范式（Austin 等人，2021年）。值得注意的是，这些离散扩散模型（离散扩散模型）如今在语言建模中展现出强劲表现，而该领域长期由自回归（AR）架构主导。

在这一领域中，DDM 研究已分化为两个截然不同的分支：掩码扩散语言模型（MDLMs）（Sahoo 等人，2024年）和均匀噪声扩散语言模型（UDLMs）（Schiff 等人，2024年）。MDLMs 在似然建模和零样本泛化方面表现出色，但在推理步数有限时，往往难以生成连贯且上下文一致的输出。相反，UDLMs 擅长少步生成，但随着推理步数增加，其表现常常落后于 MDLMs。尽管各有优势，两者在这些方面都未能实现均衡的性能。

这种不平衡具有重要的实践意义。在图像合成等领域，所需的推理步数通常远少于图像块总数。在这些少步数区间内，UDLM（及其隐式高斯对应模型（Schiff 等人，2024年；Sahoo 等人，2025年））始终优于掩码范式。具体而言，我们的实验表明，在 8 步区间生成 ImageNet-1K 图像时，UDLM 的表现比 MDLM 高出 17.6%。因此，如何兼顾 UDLMs 稳健的少步生成能力与 MDLMs 出色的语义理解能力，仍是一项亟待解决的关键挑战。

为此，我们提出混合扩散语言模型（XDLM），这是一种平衡的理论形式化框架，在均匀噪声分布与掩码噪声分布之间建立了严谨的桥梁。与 GIDD（von Rütte 等人，2025年）不同，GIDD 依赖计算开销巨大的非齐次过程，噪声特性在每个时间步都会变化；XDLM 则采用平稳噪声核，使增量噪声在结构上与边缘噪声相匹配，如图1 （左）所示。这种一致性使我们的训练目标能够高效地分解为静态常数和动态调度，从而避免 GIDD 所需的复杂噪声分布重新计算。此外，我们证明 MDLM 和 UDLM 是该形式化框架的极限情形。XDLM 还通过代数化简后验计算，实现了内存高效的实现，使其能够扩展至大词表大小，同时避免高昂的计算成本。

[1]中国科学院大学 [2]小红书。通讯作者：Yunfan Liu< liuyunfan@ucas.ac.cn>、Shaosheng Cao< caoshaosheng@xiaohongshu.com>。

预印本。*2026*年*2*月*3*日。

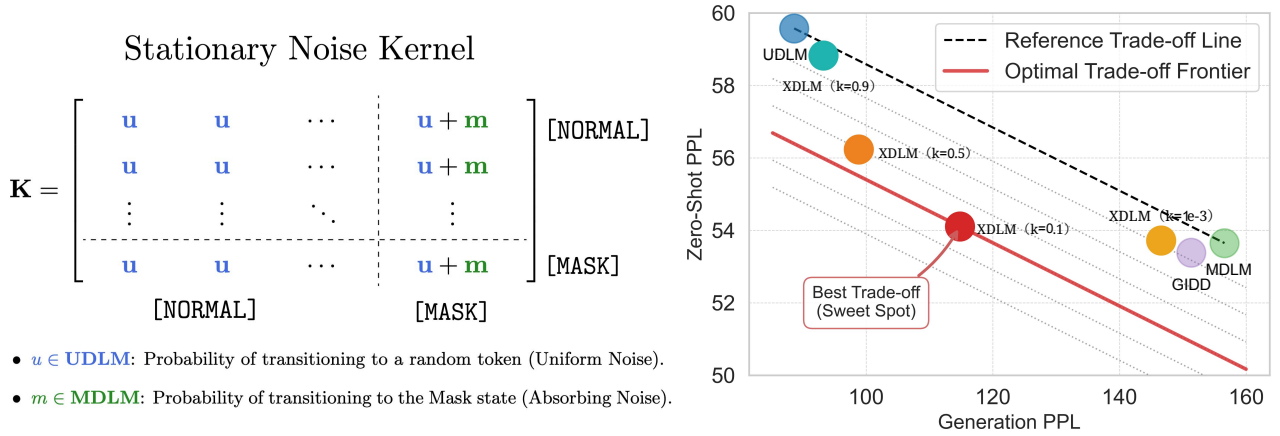

图 *1.* 左：XDLM 结合 UDLM（u）和 MDLM（m）的噪声核，在两种方法之间实现了良好的权衡。［普通词元］表示普通词元，而［掩码］表示掩码词元。右：理解能力（零样本困惑度；越低越好）与生成能力（32 个采样步数下的生成困惑度；越低越好）之间的权衡。所提出的混合比例为 $k = 0.1$ 的 XDLM 实现了最佳平衡，标记为“最佳平衡点”。

实验结果表明，XDLM 在多种模态上始终取得更优表现，验证了我们提出的平衡离散扩散方法的有效性。如图1（右）所示，XDLM 推进了理解能力与生成质量之间的帕累托前沿。值得注意的是，随着插值参数趋向极端，XDLM 的表现分别与 UDLM 和 MDLM 相当，印证了该模型的理论基础。具体而言，在零样本语言基准测试中，XDLM 的平均指标比 UDLM 高 5.4 分，仅比 MDLM 低 0.45 分。在条件图像生成任务中，XDLM 在效率和质量方面均表现出色：它在少步生成中显著优于 MDLM（4 步时 FID 从 80.8 降至 54.1），并在 16 步时排名第一，领先 UDLM 0.4 分。此外，对 8B 参数大语言模型进行大规模持续预训练（Nie 等人，2025年）后，仅用 32 个采样步数便取得 15.0 的 MBPP 分数，比原始 LLaDA 基线高出 120% 以上（15.0 对比 6.8）。最后，训练动态分析揭示了明显的性能交叉：掩码基线模型虽收敛迅速，却很早便陷入平台期；而 XDLM 在整个训练过程中持续稳步提升，展现出更强的长期扩展能力。

2. 预备知识

离散扩散模型（DDMs）是由两个马尔可夫过程定义的潜在生成模型：一个前向过程和一个反向过程。考虑单个类别变量$\mathbf{x}_0$,，它在 $\{0,1\}^N$ 中表示为独热向量，其中 $N = |\mathcal{V}|$ 表示词汇表大小。令$\mathbf{z}_t$ 表示时间 $t \in [0,1]$ 的潜在状态，其中$\mathbf{z}_0 \coloneqq \mathbf{x}_0$。前向过程 q 通过转移核 $q(\mathbf{z}_t \mid \mathbf{z}_s)$ 逐步破坏数据，其中 $s < t$，而反向过程 $p_\theta(\mathbf{z}_s \mid \mathbf{z}_t)$ 由 θ 参数化，通过迭代对潜在状态进行去噪，以恢复干净数据$\mathbf{x}_0$。

前向过程 前向转移概率由行随机矩阵 $\mathbf{Q}_{t|s} \in \mathbb{R}^{N\times N}$ 控制：

$$q(\mathbf{z}_t \mid \mathbf{z}_s) = \mathrm{Cat}(\mathbf{z}_t; \mathbf{Q}_{t|s}^\top \mathbf{z}_s), \tag{1}$$

其中 $\mathbf{z}_t$ 和$\mathbf{z}_s$ 由于潜在状态的离散性而为独热向量，而 $\mathrm{Cat}(\mathbf{x};\mathbf{p})$ 是一个类别分布。对于$\mathbf{Q}^{t|s}$，常见的结构选择包括均匀转移（UDLM（Schiff 等人，2024年））和吸收态转移（MDLM（Sahoo 等人，2024年））。

反向过程 利用贝叶斯定理和马尔可夫性质，给定$\mathbf{z}_s$ $\mathbf{z}_t$ 和$\mathbf{x}_0$ 时的后验概率可表示为（Austin 等人，2021年）：

$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}_0) = \frac{(\mathbf{Q}_{t|s}\mathbf{z}_t) \odot (\mathbf{Q}_{s|0}^\top \mathbf{x}_0)}{\mathbf{z}_t^\top \mathbf{Q}_{t|0}^\top \mathbf{x}_0}. \tag{2}$$

在文献中，该反向转移通常使用 x_0-参数化进行近似（Austin 等，2021年；Sahoo 等，2024年；Schiff 等，2024年），其中采用神经网络 f_θ 来预测干净数据 $\tilde{\mathbf{x}}_0 = f_\theta(\mathbf{z}_t)$ 的分布。因此，参数化反向转移可写为 $p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = q(\mathbf{z}_s \mid \mathbf{z}_t, \tilde{\mathbf{x}}_0)$。

训练目标 变分下界（VLB，证据下界）在训练过程中被优化，它可以分解为三个组成部分：先验项、扩散项和重构项。从数学上定义 $s = (i-1)/T$和

$t = i/T$，目标 $\mathcal{L}_{\text{vb}}$ 可以表述为 as:

$$\mathcal{L}_{\text{vb}} = \mathbb{E}_{q(\mathbf{x}_0)}\Bigg[\underbrace{D_{\text{KL}}(q(\mathbf{z}_1 \mid \mathbf{x}_0) \parallel p(\mathbf{z}_1))}_{\mathcal{L}_{\text{prior}}} + \sum_{i=2}^{T} \underbrace{\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x}_0)}[D_{\text{KL}}(q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{z}_s \mid \mathbf{z}_t))]}_{\mathcal{L}_{\text{diffusion}}} - \underbrace{\mathbb{E}_{q(\mathbf{z}_{1/T}|\mathbf{x}_0)}[\log p_\theta(\mathbf{x}_0 \mid \mathbf{z}_{1/T})]}_{\mathcal{L}_{\text{recons}}}\Bigg]. \quad (3)$$

在连续时间极限（$T \to \infty$）下，先验项和重构项均简化为 0，使目标函数简化为扩散项（Schiff 等人，2024年）：

$$\mathcal{L}_{\text{vb}} = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\left[T D_{\text{KL}}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{z}_s|\mathbf{z}_t))\right]. \quad (4)$$

3. 基于平稳噪声核的混合扩散语言建模

本节详细介绍所提出的混合扩散语言建模（XDLM）框架。首先在第3.1节介绍采用平稳噪声核的前向过程，然后在第3.2节推导用于高效采样和训练的标量形式。最后，第3.3节说明，在特定混合参数设置下，UDLM 和 MDLM 可视为 XDLM 的极限情形。

3.1. 采用平稳核的前向过程

离散扩散模型从时间 s 到 t 的前向扩散过程由转移矩阵 $\mathbf{Q}_{t|s}$ 控制，该矩阵定义为信号保持恒等矩阵$\mathbf{I}$ 与平稳噪声核$\mathbf{K}$的凸组合：

$$\mathbf{Q}_{t|s} = \alpha_{t|s}\mathbf{I} + \beta_{t|s}\mathbf{K}, \quad (5)$$

其中 $\alpha_{t|s}, \beta_{t|s} \in [0,1]$ 表示满足 $\alpha_{t|s} + \beta_{t|s} = 1$ 的标量调度。

我们方法中的一个关键设计选择是 $\mathbf{K}$ 的平稳性。不同于 GIDD（von Rütte 等人，2025年），后者利用时变混合分布，使噪声结构与扩散时间线天然耦合；我们要求 $\mathbf{K}$ 在所有时间步中保持不变。这种平稳性将噪声特性与调度动态解耦，确保扩散过程沿着几何上一致的轨迹趋向固定的目标分布，这与流匹配（Lipman 等人，2022年）提出的最优传输路径一致。

在此约束下，噪声核$\mathbf{K}$ 必须将每个输入状态映射到目标噪声分布 $\boldsymbol{\pi}$，而与起始状态无关。该要求将在下文形式化。

引理 3.1（混合核的构造）. 设 $\boldsymbol{\pi} \in \mathcal{R}^N$ 为平稳目标分布，定义为均匀分布$\mathbf{u}$ 与特殊词元 $i \in \mathcal{S}$上的点质量的混合。若噪声核$\mathbf{K}$满足瞬时混合性质（即单步收敛到 $\boldsymbol{\pi}$），则它可作如下分解：

$$\mathbf{K} = \frac{k}{N}\mathbf{J} + \sum_{i\in\mathcal{S}} \mu_i \mathbf{M}_i, \quad (6)$$

其中 $\mathbf{J}$ 是全一矩阵，$\mathbf{M}_i$ 是状态 i对应的吸收矩阵，权重满足 $k + \sum \mu_i = 1$。

我们将引理 3.1（推导见附录 A）具体化到标准的 MDLM 设定，其中只有一个掩码词元，以独热向量 $\mathbf{m}$表示。令 k 和 μ 分别表示均匀分量和掩码分量的权重，由此得到的噪声核矩阵自然地在均匀噪声与掩码状态之间进行插值：

$$\mathbf{K} = \frac{k}{N}\mathbf{J} + \mu\mathbf{M}. \quad (7)$$

图 1展示了噪声核 K 中的转移动力学，该动力学由均匀噪声和吸收噪声（掩码状态）驱动，分别标记为 $\mathbf{u}$ 和 $\mathbf{m}$。

3.2. 通过标量形式实现高效采样与训练

将噪声核 $\mathbf{K}$（式 6）和转移矩阵 $\mathbf{Q}^{t|s}$（式 5）的定义代入式 2 和式 4后，可以直接得到后验 $q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x})$ 和 KL 散度的详细形式。然而，通过矩阵运算直接计算这些表达式，在内存和计算方面的开销都高得难以承受，尤其是在词表规模较大时。在本节中，我们通过引理 3.3 和引理 3.4推导出，后验和 KL 散度可以重新表述为等价的标量形式。进一步结合引理 3.5所定义的渐近极限行为，我们建立了一个易于计算且数值稳定的训练目标，从而避免对大词汇表大小进行昂贵的显式计算。

具体而言，设$\mathbf{e}$ 表示任意一个独热基向量，则转移到状态$\mathbf{e}$ 的后验概率可写为

$$q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \mathbf{x}) = \mathbf{e}^\top \frac{(\alpha_{t|s}\mathbf{z_t} + \beta_{t|s}K\mathbf{z_t}) \odot (\alpha_s\mathbf{x} + \beta_s K^\top\mathbf{x})}{\mathbf{z_t}^\top(\alpha_t\mathbf{x} + \beta_t K^\top\mathbf{x})}, \quad (8)$$

我们观察到，式 8可通过定义一组辅助函数，改写为等价的标量形式，从而显著降低后验和 KL 散度的计算复杂度。

定义 3.2（标量原语）。令 $p_{\mathbf{x},\mathbf{e}} \coloneqq \mathbf{e}^\top \mathbf{x}$ 表示分布 $\mathbf{x}$ 在词元 $\mathbf{e}$处的概率质量。我们将噪声率函数 $r(\mathbf{e})$ 和前向扩散映射 $f_t(\mathbf{x},\mathbf{e})$ 定义为：

$$r(\mathbf{e}) = \frac{k}{N} + \mu\delta_{\mathbf{e},\mathbf{m}} \tag{9}$$

$$f_t(\mathbf{x},\mathbf{e}) = \alpha_t p_{\mathbf{x},\mathbf{e}} + \beta_t r(\mathbf{e}) \tag{10}$$

这些辅助函数为后验和 KL 散度提供了以下高效表达式，完整推导见附录 B。

引理3.3（标量后验分布）。转移到状态$\mathbf{e}$ 的后验概率可表示为：

$$q(\mathbf{z}_s = \mathbf{e} \mid \mathbf{z}_t, \mathbf{x}) = \frac{f_s(\mathbf{x},\mathbf{e}) f_{t|s}(\mathbf{e},\mathbf{z}_t)}{f_t(\mathbf{x},\mathbf{z}_t)}. \tag{11}$$

模型的反向转移 p_θ 形式相同，只需将$\mathbf{x}$ 替换为预测分布 $\tilde{\mathbf{x}}_0$。

引理3.4（标量KL散度）。项 $D_{KL}(q(\mathbf{z}_s \mid \mathbf{z}_t,\mathbf{x}) \parallel p_\theta(\mathbf{z}_s \mid \mathbf{z}_t))$ 可写为：

$$D_{KL} = \frac{\beta_{t|s}\alpha_s r(\mathbf{z}_t)}{f_t(\mathbf{x},\mathbf{z}_t)} h_t(\mathbf{x},\mathbf{z}_t,\tilde{\mathbf{x}}_0), \tag{12}$$

其中 h_t 是一个用于汇总对数差项的辅助函数：

$$\begin{aligned} h_t(\mathbf{x},\mathbf{z}_t,\tilde{\mathbf{x}}_0) = {} & \frac{f_s(\mathbf{x},\mathbf{z}_t)}{r(\mathbf{z}_t)} \frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s} \log \frac{f_s(\mathbf{x},\mathbf{z}_t) f_t(\tilde{\mathbf{x}}_0,\mathbf{z}_t)}{f_t(\mathbf{x},\mathbf{z}_t) f_s(\tilde{\mathbf{x}}_0,\mathbf{z}_t)} \\ & - \frac{1}{\alpha_s} \log \frac{f_t(\mathbf{x},\mathbf{z}_t)}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z}_t)} + \log \frac{f_s(\mathbf{x},\mathbf{x})}{f_s(\tilde{\mathbf{x}}_0,\mathbf{x})} \\ & + \frac{k\beta_s}{N\alpha_s} \sum_{\mathbf{e}\in\mathcal{V}} \log \frac{f_s(\mathbf{x},\mathbf{e})}{f_s(\tilde{\mathbf{x}}_0,\mathbf{e})}. \end{aligned} \tag{13}$$

对于逼近连续时间的极限情形 $s \to t$，辅助函数 h_t 会得到简化，从而规避与公式 13 中第一个对数项相关的数值不稳定性。

引理3.5（极限情形）。当 $s \to t$ 时，函数 h_t收敛为：

$$\begin{aligned} h_t(\mathbf{x},\mathbf{z}_t,\tilde{\mathbf{x}}_0) \approx {} & \frac{p_{\mathbf{x},\mathbf{z}_t} - p_{\tilde{\mathbf{x}}_0,\mathbf{z}_t}}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z}_t)} - \frac{1}{\alpha_t} \log \frac{f_t(\mathbf{x},\mathbf{z}_t)}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z}_t)} + \\ & \log \frac{f_t(\mathbf{x},\mathbf{x})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{x})} + \frac{k\beta_t}{N\alpha_t} \sum_{\mathbf{e}\in\mathcal{V}} \log \frac{f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})}. \end{aligned} \tag{14}$$

因此，将引理3.4 和 3.5 中推导出的渐近行为代入式 4 定义的训练目标后，我们得到训练损失的一个易处理标量形式：

$$\lim_{T\to\infty} \mathcal{L}_{\text{vb}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{z}_t} \left[\frac{-\alpha_t' r(\mathbf{z}_t)}{f_t(\mathbf{x},\mathbf{z}_t)} \cdot h_t(\mathbf{x},\mathbf{z}_t,\tilde{\mathbf{x}}_0) \right]. \tag{15}$$

这种代入方式避免了显式计算后验分布及其 KL 散度的高昂成本，从而有助于高效、稳定地实现 XDLM 的训练和采样。

3.3. 与 MDLM 和 UDLM 的关系

所提出的 XDLM 对现有的 DDM 框架进行了泛化。通过调节均匀噪声权重k （公式7），并相应调整吸收强度，XDLM 在参数取极限时可退化为 UDLM 和 MDLM。本节概述 XDLM 与这些模型之间的关系，详细分析留待附录 C。

与 MDLM 的联系。令 $k = 0$（这意味着$\mu = 1$），噪声核将退化为纯掩码操作。因此，后验概率可简化为：

$$\begin{aligned} & p_\theta(\mathbf{z}_s = \mathbf{e} \mid \mathbf{z}_t) \\ & = \delta_{\mathbf{z}_t,\mathbf{m}}\delta_{\mathbf{e},\mathbf{m}} \frac{\beta_s}{\beta_t} + \delta_{\mathbf{z}_t,\mathbf{m}}\delta_{\mathbf{e}\neq\mathbf{m}} \frac{\beta_{t|s}\alpha_s p_{\theta,\mathbf{e}}}{\beta_t} + \delta_{\mathbf{z}_t\neq\mathbf{m}}\delta_{\mathbf{e},\mathbf{z}_t} \\ & = \begin{cases} \frac{\beta_s}{\beta_t}\delta_{\mathbf{m},\mathbf{z}_t}, & \mathbf{e} = \mathbf{m} \\ \frac{(\alpha_s - \alpha_t)}{1-\alpha_t} p_{\theta,\mathbf{e}}, & \mathbf{z}_t = \mathbf{m}, \mathbf{e} \neq \mathbf{m} \\ \delta_{\mathbf{e},\mathbf{z}_t}, & \mathbf{z}_t \neq \mathbf{m}, \mathbf{e} \neq \mathbf{m} \end{cases} \end{aligned} \tag{16}$$

因此，KL 散度会退化为对掩码词元的标准交叉熵损失，与 MDLM 目标函数（Sahoo 等人，2024年）一致：

$$\mathcal{L}_{\text{KL}} = -\frac{\beta_{t|s}\alpha_s}{\beta_t} \log p_\theta(\mathbf{x} \mid \mathbf{z}_t). \tag{17}$$

与 UDLM 的联系。设定 $k = 1$ （意味着 $\mu =$为 0）可得到均匀噪声核。在此极限下，后验概率变为：

$$\begin{aligned} & p_\theta(\mathbf{z}_s = \mathbf{e} \mid \mathbf{z}_t) \\ & = \frac{\delta_{\mathbf{e},\mathbf{z}_t}(N\alpha_t p_{\theta,\mathbf{e}} + \beta_s\alpha_{t|s}) + (\alpha_s - \alpha_t)p_{\theta,\mathbf{e}} + \beta_s\beta_{t|s}/N}{N\alpha_t p_{\theta,\mathbf{z}_t} + \beta_t} \end{aligned} \tag{18}$$

定义 $\bar{\mathbf{x}}_\mathbf{i} = \mathbf{N}\mathbf{f}_\mathbf{t}(\mathbf{x},\mathbf{e}_\mathbf{i})$, $(\bar{\mathbf{x}}_\theta)_\mathbf{i} = \mathbf{N}\mathbf{f}_\mathbf{t}(\tilde{\mathbf{x}}_\mathbf{0},\mathbf{e}_\mathbf{i})$ 并考虑极限 $s \to t$，我们推导的 KL 散度恰好与（Schiff 等人，2024年）中推导的 UDLM 损失完全一致：

$$\text{KL} = \frac{-\beta_{t|s}\alpha_s}{N\alpha_t} \left(\frac{N}{(\bar{x}_\theta)_i} - \frac{N}{\bar{x}_i} - \sum_{js.t.(\mathbf{z}_t)_j=0} \frac{\bar{x}_j}{\bar{x}_i} \log \frac{(\bar{x}_\theta)_i \bar{x}_j}{(\bar{x}_\theta)_j \bar{x}_i} \right) \tag{19}$$

其中 i 是 $\mathbf{z}_t$ 中活动状态的索引。

表 *1.* OWT 数据集上的验证困惑度。在相同训练步数下，XDLM 的性能与 MDLM 相当，且优于 UDLM。

模型	MDLM	GIDD	XDLM	UDLM
PPL	23.321	23.136	24.097	25.937

4. 实验

实验设置首先在第4.1节介绍，更多细节见附录 F。第4.2节随后报告了 XDLM 在零样本似然、语言生成质量和图像生成任务上的性能，展示了其相较于现有方法的优越性，更多结果见附录G、H 和I。此外，为评估其扩展到更大模型规模的能力，通过在 LLaDA 上进行持续预训练，XDLM 进一步扩展为 LLaDA-XDLM，并在标准基准测试中取得了强劲的性能（更多细节见附录J）。最后，第 4.3节展示了跨不同任务的分析与可视化结果，表明 XDLM 相较于其他方法，在理解与生成之间实现了更好的平衡（更多结果见附录K、L、D和E）。

4.1. 实验设置

所有实验均在由 8× 台 H800 GPU 组成的集群上进行。在适用情况下，我们统一了关键超参数，采用 AdamW 优化器（$\beta_1 = 0.9$，$\beta_2 = 0.999$，权重衰减为 0）、$k = 0.1$ 的混合比例、512 的全局批次大小，以及 0.9999 的 EMA 率。

4.2. 性能

零样本似然评估。为评估 XDLM 的泛化能力，我们在 OpenWebText（OWT）上完成训练后，在七个外部数据集上评估零样本困惑度（PPL）。与先前工作（Sahoo 等人，2024年；2025年）保持一致，我们的评估套件包括 AG News（Zhang 等人，2015）、LAM-BADA（Paperno 等人，2016）、LM1B（Chelba 等人，2013年）、宾州树库（Marcus 等人，1993）、来自 arXiv 和 PubMed 的科学论文（Cohan 等人，2018）以及 WikiText（Merity 等人，2016）的验证集划分。验证困惑度通过结合蒙特卡洛采样的负证据下界（ELBO）进行估计。为确保公平比较，我们采用了（Sahoo 等人，2024年；2025年）中完全相同的采样配置。

如表 1所示，XDLM 在 OWT 验证集上的困惑度为 24.097，与 MDLM 和 GIDD 相当，并显著优于 UDLM（25.937）。这种对域内能力的保留也延伸到了零样本基准测试。如表 2所列，XDLM 在七个数据集上的平均困惑度为 54.110，表现与.....相近 MDLM（53.650）和 GIDD（53.384），而 UDLM 明显落后（59.574）。

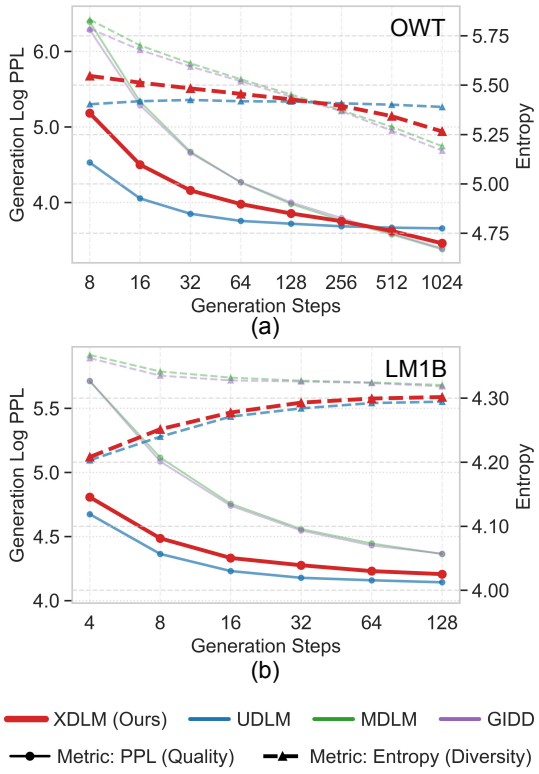

图 *2.* 语言生成质量。OWT（上）和 LM1B（下）的结果表明，XDLM 在少步数和多步数区间均实现了更好的平衡。为便于清晰呈现，困惑度和生成步数均采用对数刻度。

语言生成质量。我们使用困惑度（质量）和词元熵（多样性）评估祖先采样（Sahoo 等人，2024年；Schiff 等人，2024年；Sahoo 等人，2025年）。图2显示，XDLM 在掩码范式与均匀扩散范式之间实现了稳健的平衡。在少步数区间（OWT 上为 8–32 步），XDLM 得益于均匀噪声动态，生成质量可与 UDLM 相媲美，并明显优于纯掩码模型。相比之下，在多步数区间（512–1024 步），掩码噪声占据主导，使 XDLM 优于 UDLM，达到 MDLM 水平的性能。在 LM1B 上也观察到类似趋势，证实 XDLM 能够在保留掩码严谨结构的同时，确保均匀噪声的效率。关于混合比例影响的全面数值分析 k 见附录H。

图像生成性能。我们在 ImageNet-1K 上通过类别条件生成评估领域泛化能力。表 3 报告了未使用无分类器引导（CFG）时的 FID 和 IS（初始分数）。在少步数区间，XDLM 和 UDLM 始终优于基于掩码的基线方法（即 MDLM 和 GIDD）。XDLM 表现出与 UDLM 极强的竞争力，在 16 步时取得最低 FID（25.77），展现了其高效生成高质量内容的能力。当

表 2. 各项基准测试上的零样本性能比较。XDLM 的性能与 MDLM 和 GIDD 持平，显著优于 UDLM。各模型均在统一设置下使用 8× H800 GPU 重新实现。

数据集	AG News	LAMBADA	LM1B-GPT2	PTB	arXiv	PubMed	WikiText	平均值
MDLM	61.374	47.967	65.629	89.049	37.457	41.981	32.093	53.650
GIDD	60.607	47.811	65.898	86.911	39.019	42.634	30.809	53.384
XDLM	62.768	45.608	68.229	90.796	37.232	41.391	32.748	54.110
UDLM	69.402	51.272	75.572	95.986	42.671	47.181	34.933	59.574

表 3. 采用标准条件时，ImageNet-1K 上的图像生成性能比较。

模型	FID ↓ 4步	8步	16步	IS ↑ 4步	8步	16步
MDLM	80.752	47.732	28.785	16.287	29.178	44.656
GIDD	86.842	54.933	35.403	14.559	24.297	35.698
XDLM	5					
UDLM	49.861	30.144	26.242	27.049	38.832	41.801

表 4. 使用 CFG $scale = 2.0$ 的 ImageNet-1K 图像生成性能比较。

模型	FID ↓ 4步	8步	16步	IS ↑ 4步	8步	16步
MDLM	33.468	11.144	6.725	54.740	119.150	172.664
GIDD	41.000	15.151	7.076	44.084	95.789	148.148
XDLM	13.550	8.956	8.625	107.403		
UDLM	14.055	9.718	8.980	97.859	123.582	132.099

使用 CFG 进行评估（尺度为=2.0），如表 4所示，纯基于掩码的 MDLM 在 16 步时取得了最低的总体 FID（6.73），这表明在步数充足时，掩码方法表现出色。然而，由于部分继承了吸收噪声的特性，XDLM 在少步数区间显著优于 UDLM，在 4 步（13.55）和 8 步（8.96）时取得最佳 FID 分数。更详细的分析见附录I。

大语言模型的持续预训练（LLMs）。为了验证 XDLM 能否作为最先进大语言模型的可扩展升级方案，我们将预训练的 8B-MDLM LLaDA（Nie 等人，2025年）通过计算高效的 600 步预训练（称为 LLaDA-XDLM）改造为 XDLM形式。为严格区分性能提升的来源，我们将其与两种管控措施进行比较：LLaDA-MDLM（相同持续时间的标准持续预训练）和 LLaDA-XDLM-infer（将 LLaDA 与我们的采样策略结合）。

在 OpenCompass 上的评测（32 步）显示，LLaDA-XDLM 在 GSM8K、MATH 和 BBH 上均持续优于原始 LLaDA 及两种管控措施（图3 （a））。这种影响在 MBPP 代码生成任务中最为明显：LLaDA-XDLM 将 LLaDA 的性能提升了一倍（15.0对比6.8），同时减少了无法编译的错误，从而证实其结构连贯性得到增强。

至于对照模型，LLaDA-MDLM 的表现较原始 LLaDA 有所下降，这很可能是由于微调语料库与预训练语料库之间存在分布偏移；而在相同条件下，LLaDA-XDLM 则取得了显著提升。这证实了性能增益源于 XDLM 形式的有效性，而非仅仅来自额外的 600 个训练步数。此外，LLaDA-XDLM 的表现显著优于 LLaDA-XDLM-infer，这表明改进是模型学习到的内在特性，而非采样策略造成的假象。

4.3. 分析

训练动态中的性能交叉。对训练动态的分析揭示了一种“性能交叉”现象：随着训练推进，竞争模型的相对效能会发生显著变化。LM1B 文本生成任务尤能体现这一趋势，如图4 (a) 中的生成困惑度曲线所示。虽然 GIDD 和 MDLM 在初始阶段（<200k 步，粉色区域）表现出优势，但它们很快便趋于饱和，最终被 UDLM 超越（黄色区域）。凭借从 UDLM 继承的均匀特性，XDLM 在后期取得了显著提升，经过 700k 步后最终达到或超过 MDLM（绿色区域）。这一发展轨迹与 GIDD 截然不同：GIDD 早早便趋于饱和，在整个训练过程中困惑度仅有微幅改善。

ImageNet-1K 图像生成中也呈现出另一种趋势迥异但相似的动态，如图4 (b) 所示。从一开始，XDLM 就表现出更优的性能，始终取得最低的 FID 分数。UDLM 虽然最终在训练最后阶段（> 500k 步）与 XDLM 收敛至相当水平，但在整个过程中，这两个模型的表现都显著优于 MDLM 和 GIDD 基线。

混合比例的影响。混合比例 k 是一个关键控制参数，可在两种截然不同的范式之间进行插值：UDLM（$k \to 1$）和 MDLM（$k \to 0$），二者在生成和理解方面的能力各不相同。如图1 （右）所示，这两个边界情况确定了一条线性参考权衡线（黑色虚线），代表一种零和情形：生成质量的提升（生成 PPL 更低）是以理解能力下降（零样本困惑度更高）为代价的。

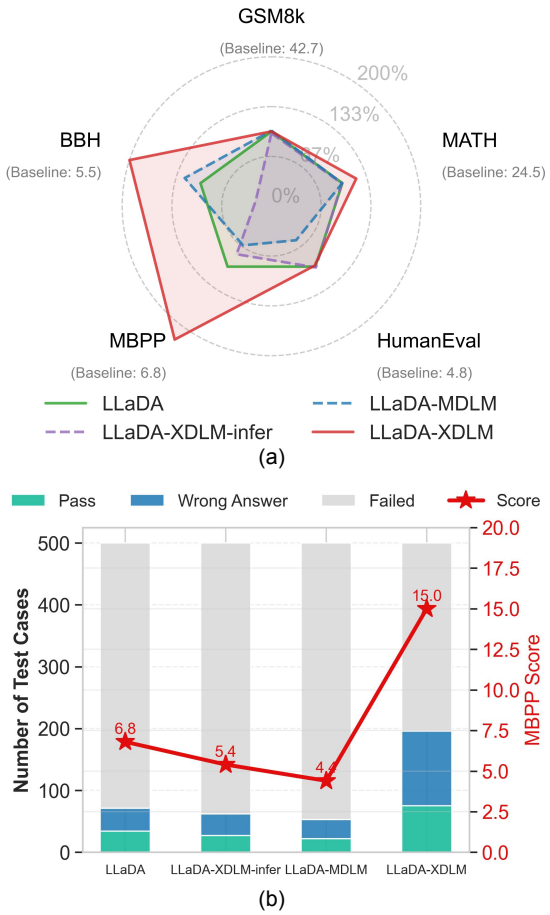

图3. 将 LLaDA-8B 适配到我们的 XDLM 形式化（LLaDA-XDLM）的评估：(a) LLaDA-XDLM 在各类基准测试中持续优于基线；(b) 改进在代码生成（MBPP）中尤为显著，该模型大幅减少了生成失败。

然而，XDLM 的实证表现表明，模型并非只沿着这条线性路径移动。关键在于，改变 k 会使模型的运行点沿着一条连续、平滑的轨迹移动。这种连贯的演变，而非随机波动或随机游走，验证了该架构设计，证实 XDLM 成功地将 UDLM 和 MDLM 的机制特性融合到一个统一框架中。此外，这条轨迹向内弯曲，形成一个更优的最优权衡前沿（红色实线），表明混合目标产生了协同效应，其性能提升超过各部分之和。沿着这一前沿探索，我们找到了一个实证上的“最佳平衡点”，位于 $k = 0.1$。如红色标记所示，这一配置有效突破了基线的帕累托前沿，在避免纯掩码模型通常导致的生成性能严重下降的同时，实现了稳健的理解能力。

XDLM 的计算效率。我们分析其在推理、训练和采样方面的计算效率。尽管 MDLM 凭借简单的吸收核仍是速度最快的基线模型，但在采用均匀噪声的模型中，XDLM 的效率最高。得益于我们重新设计的标量策略，XDLM 的前向吞吐量达到每秒 396,398 个词元，接近 GIDD 每秒 199,516 个词元的两倍；训练吞吐量则达到每秒 137,372 个词元。在生成任务中，XDLM 的采样速度达到每秒 7,108 个词元，显著超过 UDLM（每秒 2,882 个词元）。其高效率也体现在内存消耗上：XDLM 仅需 31.4 GB，而 UDLM 和 GIDD 分别需要 59.7 GB 和 40.9 GB。这些结果证实，XDLM 能有效缓解大词表带来的瓶颈，同时不牺牲复杂噪声核的表达能力。更多详细结果请参阅附录K。

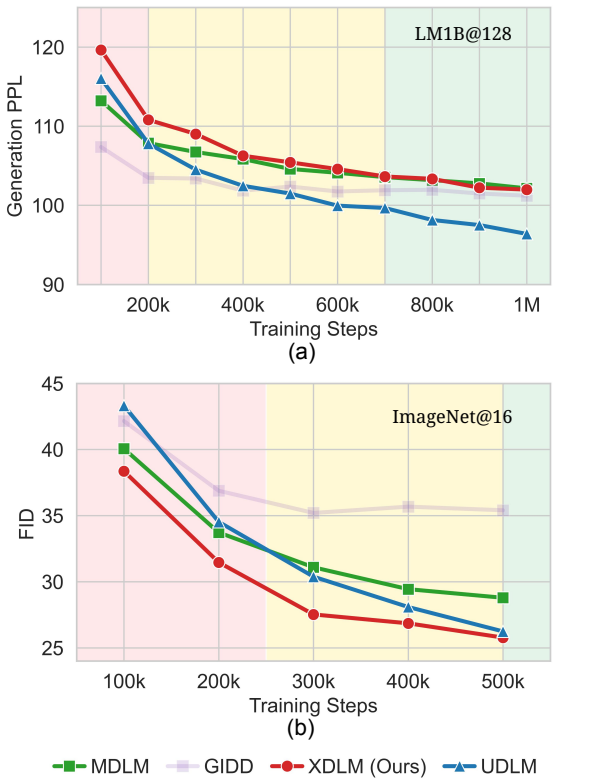

图4. （a）文本与（b）图像生成任务的训练动态。彩色区域表示各阶段中占主导地位的模型，而颜色之间的过渡则标示性能交叉点。

推理动态。为了探究 XDLM 的内部生成机制，我们仔细检查了一个长度为 $T = 32$ 步的生成轨迹。结果表明，XDLM 有效利用了一种混合噪声过程：通过吸收噪声（[MASK] $\rightarrow$ 词元）构建语义结构，同时通过均匀噪声（词元 $\rightarrow$词元）不断优化内容。关键的是，XDLM 凭借重新掩码策略（将词元转换回 [MASK]）展现出独特之处。与那些对初始化噪声高度敏感的标准迭代模型不同，

这种机制使 XDLM 能够主动舍弃由均匀噪声产生的低概率词元，从而有效摆脱局部最优。详细分析请参阅附录D 和E。

5. 相关工作

掩码扩散语言建模。在 D3PM（Austin 等人，2021年）提出的通用且灵活的离散扩散框架基础上，基于吸收态的腐蚀（通常使用特殊的 [MASK] 词元）已成为语言离散扩散中的一种流行范式，常被称为掩码扩散语言模型（MDLMs）。MDLM（Sahoo 等，2024）和 MD4（Shi等人，2024年）扩展了该框架，并将掩码扩散的连续时间变分目标推导为交叉熵损失的简单加权积分，从而为离散语言模型提供了精简且通用的训练方案。从连续时间视角来看，CTMC（Campbell等人，2022年）通过连续时间马尔可夫链形式化了前向过程，而 SEDD（Lou 等，2023）直接学习密度比，提供了分数匹配的离散类比。这些进展推动了掩码扩散语言模型向更大模型规模和数据集扩展，催生了 LLaDA（Nie 等，2025）和 Dream（Ye等人，2025年）等大型扩散语言模型，并进一步扩展到多模态设置（You等人，2025年；Yang等人，2025年；Yu等人，2025年）。

均匀噪声扩散语言建模。均匀噪声扩散的转移核是恒等变换与词表上的均匀分布的混合。该思想最初由 Argmax/Multinomial 扩散（Hoogeboom 等人，2021年）针对离散空间提出，随后由 D3PM（Austin 等人，2021年）进行了推广。Plaid（Gulrajani与Hashimoto，2023年）在缩小自回归模型与基于扩散的语言模型之间的似然差距方面迈出了重要一步。UDLM（Schiff 等人，2024年）表明，借助恰当的引导机制，均匀噪声扩散能够实现更强的可控性。Duo（Sahoo 等人，2025年）揭示了高斯扩散语言模型（如 Plaid）与均匀噪声扩散之间的对偶性，并提出从辅助潜在高斯模型中蒸馏出均匀噪声模型。

混合语言建模。D3PM（Austin 等人，2021年）指出，BERT 可以被视为一种带有转移矩阵的单步混合腐蚀实例。然而，BERT 仅针对被掩码的位置采用交叉熵目标进行训练，而基于扩散的方法则优化在整个噪声调度（或时间）上积分的目标，因此会产生不同的训练动态和归纳偏置。

近期研究进一步融合并推广了这些范式。（Fathi 等人，2025年）通过为不同词元位置分配不同的噪声调度，统一了自回归模型与扩散模型。类似地，ReMDM（Wang 等人，2025年）放宽了前向过程严格的吸收约束，使 [掩码] 词元能够恢复为其原始状态，从而支持即插即用的编辑功能。

与我们最相关的工作是 GIDD（von Rütte 等人，2025年），该工作同样探索了掩码扩散与均匀噪声扩散之间的插值。不过，GIDD 通常依赖于构造转移矩阵，动态融合这两种机制。相比之下，我们的 XDLM 框架采用平稳噪声核。这一设计选择明确地将噪声特性与时间动态解耦，使我们能够推导出一种高效的标量形式，从而简化采样和训练。

6. 结论

本文提出了 XDLM，这是一种统一方法，从理论上弥合了掩码扩散与均匀噪声扩散之间的差距。通过将前向过程重新定义为加权的行随机转移，我们证明了 XDLM 能够将现有范式（MDLM 和 UDLM）作为特例加以还原。为确保其实用性，我们推导出一种节省内存的实现方式，降低了计算复杂度，使其能够在大型词表上进行训练。

从实证结果来看，XDLM 打破了理解与生成之间的帕累托前沿。我们确定了$k = 0.1$ 这一混合比例是最佳“平衡点”：在此比例下，模型兼具掩码模型稳健的零样本似然，以及均匀噪声模型更出色的采样多样性和少步生成质量。这些优势还体现在多个领域：XDLM 在 ImageNet-1K 上达到最先进水平，并在 8B 参数大语言模型的持续预训练中展现出显著的可扩展性，其代码生成基准测试上的表现提升了一倍。

局限性 尽管取得了这些成果，未来仍有若干值得研究的方向。首先，我们尚未从头开始大规模训练 XDLM；此类预训练很可能有助于更全面地探索模型涌现出的特性。其次，我们尚未充分研究“性能交叉”现象：在涉及大量采样步数（接近自回归解码）的生成任务中，UDLM 和 XDLM 似乎优于 MDLM。第三，尚未针对语言建模和图像生成优化 XDLM 的领域专用采样策略。此外，虽然我们确认 XDLM 能够兼顾理解与生成，但在单一统一模型中，文本与视觉模态之间的相互作用及平衡仍有待研究。最后，XDLM 的后训练方案和推理加速技术仍有待未来探索。

影响声明

本研究的首要目标是推动机器学习技术达到最先进水平。因此，本研究也涉及与语言模型和基础模型这一广泛领域中同类研究相似的社会层面考量。

参考文献

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. 离散状态空间中的结构化去噪扩散模型。*NeurIPS*, 34:17981–17993, 2021年。

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. 离散去噪模型的连续时间框架。*NeurIPS*, 35: 28266–28279, 2022年。Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. 用于衡量统计语言建模进展的十亿词基准测试。*arXiv*预印本 *arXiv:1312.3005*, 2013年。Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. 用于长文档抽象式摘要的话语感知注意力模型。*arXiv*预印本 *arXiv:1804.05685*, 2018年。Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet：大型分层图像数据库。载于*IEEE*计算机视觉与模式识别会议, 第248–255页。IEEE, 2009年。Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT：用于语言理解的深度双向变换器预训练。载于北美计算语言学协会年会, 第4171–4186页, 2019年。Dhariwal, P. and Nichol, A. 扩散模型在图像合成方面胜过生成对抗网络。*NeurIPS*, 34:8780–8794, 2021年。Fathi, N., Scholak, T., and Noël, P.-A. 统一自回归与基于扩散的序列生成。*arXiv*预印本 *arXiv:2504.06416*, 2025年。Gokaslan, A. and Cohen, V. OpenWebText语料库。
h
t
t
p://
Skylion007.github.io/OpenWebTextCorpus, 2019年。Gulrajani, I. and Hashimoto, T. 基于似然的扩散语言模型。*NeurIPS*, 36:16693–16715, 2023年。Ho, J., Jain, A., and Abbeel, P. 去噪扩散概率模型。*NeurIPS*, 33:6840–6851, 2020年。Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., and Welling, M. Argmax流与多项式扩散：学习分类分布。*NeurIPS*, 34:12454–12465, 2021年。

Kong, Z., Ping, W., Huang, J., Zhao, K., 和 Catanzaro, B. Diffwave：一种用于音频合成的通用扩散模型。*arXiv*预印本 *arXiv:2009.09761*，2020年。Krizhevsky, A., Hinton, G., 等. 从微型图像中学习多层特征。2009年。Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., 和 Le, M. 用于生成式建模的流匹配。*arXiv*预印本 *arXiv:2210.02747*, 2022年。Lou, A., Meng, C., 和 Ermon, S. 通过估计数据分布的比率进行离散扩散语言建模。2023年。Marcus, M., Santorini, B., 和 Marcinkiewicz, M. A. 构建大型英文标注语料库：宾州树库。计算语言学, 19(2):313–330，1993年。Merity, S., Xiong, C., Bradbury, J., 和 Socher, R. 指针哨兵混合模型。*arXiv*预印本 *arXiv:1609.07843*，2016年。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年。Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N.-Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., 和 Fernández, R. LAMBADA 数据集：需要广泛篇章语境的词语预测。发表于*ACL*，第1525–1534页，2016年。Peebles, W. 和 Xie, S. 使用 Transformer 的可扩展扩散模型。发表于*IEEE ICCV*，第4195–4205页，2023年。Penedo, G., Kydlíˇcek, H., Lozhkov, A., Mitchell, M., Raf- fel, C. A., Von Werra, L., Wolf, T., 等. FineWeb数据集：大规模提炼网络数据，获取优质文本数据。*NeurIPS*, 37:30811–30849，2024年。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 等. 语言模型是无监督多任务学习器。*OpenAI*博客, 1(8):9，2019年。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 和 Ommer, B. 使用潜在扩散模型进行高分辨率图像合成。发表于*IEEE*计算机视觉与模式识别会议，第10684–10695页，2022年。Ronneberger, O., Fischer, P., 和 Brox, T. U-Net：用于生物医学图像分割的卷积网络。发表于*MICCAI*，第234–241页。Springer，2015年。Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., 和 Kuleshov, V. 简单有效的掩码扩散语言模型。*NeurIPS*，37:130136–130184, 2024年。

Sahoo, S. S., Deschenaux, J., Gokaslan, A., Wang, G., Chiu, J., 和 Kuleshov, V. 扩散二元性。*arXiv*预印本 *arXiv:2506.10892*，2025年。Schiff, Y., Sahoo, S. S., Phung, H., Wang, G., Boshar, S., Dalla-torre, H., de Almeida, B. P., Rush, A. M., PIER- ROT, T., 和 Kuleshov, V. 离散扩散模型的简单引导机制。发表于*ICLR*，2024年。Shi, J., Han, K., Wang, Z., Doucet, A., 和 Titsias, M. 离散数据的简化广义掩码扩散。*NeurIPS*，37:103131–103167，2024年。Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., 和 Yuan, Z. 自回归模型胜过扩散模型：用于可扩展图像生成的 Llama。*arXiv*预印本 *arXiv:2406.06525*，2024年。von Rütte, D., Fluri, J., Ding, Y., Orvieto, A., Schölkopf, B., 和 Hofmann, T. 广义插值离散扩散。*arXiv*预印本 *arXiv:2503.04482*，2025年。Wang, G., Schiff, Y., Sahoo, S. S., 和 Kuleshov, V. 使用推理时扩展重新掩码离散扩散模型。*arXiv*预印本 *arXiv:2503.00307*，2025年。Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., 和 Wang, M. Mmada：多模态大型扩散语言模型。*arXiv*预印本 *arXiv:2505.15809*，2025年。Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., 和 Kong, L. Dream 7b：扩散大型语言模型。*arXiv*预印本 *arXiv:2508.15487*，2025年。You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.-R., 和 Li, C. Llada-v：结合视觉指令微调的大语言扩散模型。*arXiv*预印本 *arXiv:2505.16933*，2025年。Yu, R., Ma, X., 和 Wang, X. Dimple：采用并行解码的离散扩散多模态大型语言模型。*arXiv*预印本 *arXiv:2505.16990*，2025年。Zhang, X., Zhao, J., 和 LeCun, Y. 用于文本分类的字符级卷积网络。*NeurIPS*，28，2015年。

A. 具有平稳核的前向过程

引理 1（混合核的构造）。 设 $\boldsymbol{\pi} \in \mathcal{R}^N$ 为平稳目标分布，定义为均匀分布 $\mathbf{u}$ 与特殊词元 $i \in \mathcal{S}$上的点质量的混合。若噪声核 $\mathbf{K}$ 满足瞬时混合性质（即一步收敛到 $\boldsymbol{\pi}$），则它可分解为：

$$\mathbf{K} = \frac{k}{N}\mathbf{J} + \sum_{i\in\mathcal{S}} \mu_i \mathbf{M}_i, \tag{20}$$

其中 $\mathbf{J}$ 为全一矩阵，$\mathbf{M}_i$ 为状态 i 的吸收矩阵，且权重满足 $k + \sum \mu_i = 1$。

证明。 瞬时混合条件意味着转移到目标状态 j 的转移概率仅由目标边缘分布 π_j决定，而与前态 s无关。就随机矩阵而言，这要求行等价：$\mathbf{K}_{sj} = \pi_j$ 对所有 s成立。这在代数上等价于秩一分解 $\mathbf{K} = \mathbf{1}\boldsymbol{\pi}^\top$。

将混合定义 $\boldsymbol{\pi} = k\mathbf{u} + \sum_{i\in\mathcal{S}} \mu_i \mathbf{e}_i$ 代入该分解可得：

$$\mathbf{K} = \mathbf{1}\left(k\mathbf{u} + \sum_{i\in\mathcal{S}} \mu_i \mathbf{e}_i\right)^\top \tag{21}$$

$$= k(\mathbf{1}\mathbf{u}^\top) + \sum_{i\in\mathcal{S}} \mu_i (\mathbf{1}\mathbf{e}_i^\top). \tag{22}$$

将所得外积与典范转移矩阵 $\mathbf{1}\mathbf{u}^\top = N^{-1}\mathbf{J}$（均匀转移）和$\mathbf{1}\mathbf{e}_i^\top = \mathbf{M}_i$（吸收转移）对应起来，即完成证明。

□

B. 通过标量形式实现高效采样和训练

在证明引理之前，我们先介绍以下定义：

定义 B.1。

$$\mathbf{g}(s,t,\mathbf{x},\mathbf{z_t}) \tag{23}$$

$$= \alpha_t \mathbf{z_t} \odot \mathbf{x} + \alpha_{t|s}\beta_s \mathbf{z_t} \odot K^\top \mathbf{x} + \beta_{t|s}\alpha_s K\mathbf{z_t} \odot \mathbf{x} + \beta_{t|s}\beta_s K\mathbf{z_t} \odot K^\top \mathbf{x} \tag{24}$$

$$= \alpha_{t|s} f_s(\mathbf{x},\mathbf{z_t})\mathbf{z_t} + \beta_{t|s} r(\mathbf{z_t}) \left(\alpha_s \mathbf{x} + \beta_s \left(\frac{k}{N}\mathbf{1} + \mu\mathbf{m}\right)\right) \tag{25}$$

因此，我们有

$$\mathbf{e}^\top \mathbf{g}(s,t,\mathbf{x},\mathbf{z_t}) \tag{26}$$

$$= \alpha_{t|s} f_s(\mathbf{x},\mathbf{z_t})\mathbf{e}^\top \mathbf{z_t} + \beta_{t|s} r(\mathbf{z_t}) \left(\alpha_s \mathbf{e}^\top \mathbf{x} + \beta_s \left(\frac{k}{N}\mathbf{e}^\top \mathbf{1} + \mu \mathbf{e}^\top \mathbf{m}\right)\right) \tag{27}$$

$$= f_s(\mathbf{x},\mathbf{z_t})\alpha_{t|s}\mathbf{e}^\top \mathbf{z_t} + \beta_{t|s} r(\mathbf{z_t}) \left(\alpha_s \mathbf{e}^\top \mathbf{x} + \beta_s r(\mathbf{e})\right) \tag{28}$$

$$= f_s(\mathbf{x},\mathbf{e})^{\alpha_{t|s}} \mathbf{e}^\top \mathbf{z_t} + \beta_{t|s} r(\mathbf{z_t}) f_s(\mathbf{x},\mathbf{e}) \tag{29}$$

$$= f_s(\mathbf{x},\mathbf{e}) f_{t|s}(\mathbf{e},\mathbf{z_t}) \tag{30}$$

引理 2（标量后验分布）。 转移到状态$\mathbf{e}$ 的后验概率可以表述为：

$$q(\mathbf{z}_s = \mathbf{e} \mid \mathbf{z}_t, \mathbf{x}) = \frac{f_s(\mathbf{x},\mathbf{e}) f_{t|s}(\mathbf{e},\mathbf{z}_t)}{f_t(\mathbf{x},\mathbf{z}_t)}. \tag{31}$$

模型的 r　　逆向转移 p_θ 通过将$\mathbf{x}$ 替换为预测分布 $\tilde{\mathbf{x}}_0$，遵循相同的形式。

证明。

$$q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \mathbf{x}) \tag{32}$$
$$= \mathbf{e}^\top \frac{(Q_{t|s}\mathbf{z_t}) \odot (Q_s^\top \mathbf{x})}{\mathbf{z_t}^\top Q_t^\top \mathbf{x}} \tag{33}$$
$$= \mathbf{e}^\top \frac{(\alpha_{t|s}\mathbf{z_t} + \beta_{t|s}K\mathbf{z_t}) \odot (\alpha_s \mathbf{x} + \beta_s K^\top \mathbf{x})}{\mathbf{z_t}^\top (\alpha_t \mathbf{x} + \beta_t K^\top \mathbf{x})} \tag{34}$$
$$= \mathbf{e}^\top \frac{\alpha_t \mathbf{z_t} \odot \mathbf{x} + \alpha_{t|s}\beta_s \mathbf{z_t} \odot K^\top \mathbf{x} + \beta_{t|s}\alpha_s K\mathbf{z_t} \odot \mathbf{x} + \beta_{t|s}\beta_s K\mathbf{z_t} \odot K^\top \mathbf{x}}{\alpha_t \mathbf{z_t}^\top \mathbf{x} + \beta_t \mathbf{z_t}^\top K^\top \mathbf{x}} \tag{35}$$
$$= \mathbf{e}^\top \frac{\mathbf{g}(s, t, \mathbf{x}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \tag{36}$$
$$= \frac{f_s(\mathbf{x}, \mathbf{e}) f_{t|s}(\mathbf{e}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \tag{37}$$

□

引理3（标量KL散度）。项 $D_{KL}(q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) \parallel p_\theta(\mathbf{z}_s \mid \mathbf{z}_t))$ 可写为：

$$D_{KL} = \frac{\beta_{t|s}\alpha_s r(\mathbf{z}_t)}{f_t(\mathbf{x}, \mathbf{z}_t)} h_t(\mathbf{x}, \mathbf{z}_t, \tilde{\mathbf{x}}_0), \tag{38}$$

其中 h_t 是一个收集对数差项的辅助函数：

$$\begin{aligned} h_t(\mathbf{x}, \mathbf{z}_t, \tilde{\mathbf{x}}_0) = &\frac{f_s(\mathbf{x}, \mathbf{z}_t)}{r(\mathbf{z}_t)} \frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s} \log \frac{f_s(\mathbf{x}, \mathbf{z}_t) f_t(\tilde{\mathbf{x}}_0, \mathbf{z}_t)}{f_t(\mathbf{x}, \mathbf{z}_t) f_s(\tilde{\mathbf{x}}_0, \mathbf{z}_t)} \\ &- \frac{1}{\alpha_s} \log \frac{f_t(\mathbf{x}, \mathbf{z}_t)}{f_t(\tilde{\mathbf{x}}_0, \mathbf{z}_t)} + \log \frac{f_s(\mathbf{x}, \mathbf{x})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{x})} \\ &+ \frac{k\beta_s}{N\alpha_s} \sum_{\mathbf{e} \in \mathcal{V}} \log \frac{f_s(\mathbf{x}, \mathbf{e})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{e})}. \end{aligned} \tag{39}$$

证明。

$$D_{KL} = \sum_{\mathbf{e}} q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \mathbf{x}) \log \frac{q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \mathbf{x})}{p_\theta(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t})} \tag{40}$$

$$= \sum_{\mathbf{e}} q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \mathbf{x}) \log \frac{q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \mathbf{x})}{q(\mathbf{z_s} = \mathbf{e}|\mathbf{z_t}, \tilde{\mathbf{x}}_0)} \tag{41}$$

$$= \sum_{\mathbf{e}} \frac{f_s(\mathbf{x}, \mathbf{e}) f_{t|s}(\mathbf{e}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \left(\log \frac{f_s(\mathbf{x}, \mathbf{e}) f_{t|s}(\mathbf{e}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} - \log \frac{f_s(\tilde{\mathbf{x}}_0, \mathbf{e}) f_{t|s}(\mathbf{e}, \mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})} \right) \tag{42}$$

$$= \frac{1}{f_t(\mathbf{x}, \mathbf{z_t})} \sum_{\mathbf{e}} f_s(\mathbf{x}, \mathbf{e}) f_{t|s}(\mathbf{e}, \mathbf{z_t}) \log \frac{f_s(\mathbf{x}, \mathbf{e}) f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t}) f_s(\tilde{\mathbf{x}}_0, \mathbf{e})} \tag{43}$$

$$= \frac{1}{f_t(\mathbf{x}, \mathbf{z_t})} \sum_{\mathbf{e}} f_s(\mathbf{x}, \mathbf{e}) \left(\alpha_{t|s} \delta_{\mathbf{e}, \mathbf{z_t}} + \beta_{t|s} r(\mathbf{z_t}) \right) \log \frac{f_s(\mathbf{x}, \mathbf{e}) f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t}) f_s(\tilde{\mathbf{x}}_0, \mathbf{e})} \tag{44}$$

$$= \frac{\alpha_{t|s} f_s(\mathbf{x}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_s(\mathbf{x}, \mathbf{z_t}) f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t}) f_s(\tilde{\mathbf{x}}_0, \mathbf{z_t})} \tag{45}$$

$$+ \frac{\beta_{t|s} r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \sum_{\mathbf{e}} f_s(\mathbf{x}, \mathbf{e}) \tag{46}$$

$$+ \frac{\beta_{t|s} r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \sum_{\mathbf{e}} f_s(\mathbf{x}, \mathbf{e}) \log \frac{f_s(\mathbf{x}, \mathbf{e})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{e})} \tag{47}$$

$$= \frac{\alpha_{t|s} f_s(\mathbf{x}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_s(\mathbf{x}, \mathbf{z_t}) f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t}) f_s(\tilde{\mathbf{x}}_0, \mathbf{z_t})} - \frac{\beta_{t|s} r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_t(\mathbf{x}, \mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})} \tag{48}$$

$$+ \frac{\beta_{t|s} r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \sum_{\mathbf{e}} (\alpha_s \delta_{\mathbf{x}, \mathbf{e}} + \beta_s r(\mathbf{e})) \log \frac{f_s(\mathbf{x}, \mathbf{e})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{e})} \tag{49}$$

$$= \frac{\alpha_{t|s} f_s(\mathbf{x}, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_s(\mathbf{x}, \mathbf{z_t}) f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t}) f_s(\tilde{\mathbf{x}}_0, \mathbf{z_t})} - \frac{\beta_{t|s} r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_t(\mathbf{x}, \mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0, \mathbf{z_t})} \tag{50}$$

$$+ \frac{\beta_{t|s} \alpha_s r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_s(\mathbf{x}, \mathbf{x})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{x})} \tag{51}$$

$$+ \frac{k \beta_{t|s} \beta_s r(\mathbf{z_t})}{N f_t(\mathbf{x}, \mathbf{z_t})} \sum_{\mathbf{e}} \log \frac{f_s(\mathbf{x}, \mathbf{e})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{e})} + \frac{\mu \beta_{t|s} \beta_s r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} \log \frac{f_s(\mathbf{x}, \mathbf{m})}{f_s(\tilde{\mathbf{x}}_0, \mathbf{m})} \tag{52}$$

$$= \frac{\beta_{t|s} \alpha_s r(\mathbf{z_t})}{f_t(\mathbf{x}, \mathbf{z_t})} h_t(\mathbf{x}, \mathbf{z_t}) \tag{53}$$

我们省略项 $\log f_s(\mathbf{x},\mathbf{m})/f_s(\tilde{\mathbf{x}}_0,\mathbf{m})$，因为对于所有$\mathbf{x}$和 $\tilde{\mathbf{x}}_0$，该项都等于零。

□

引理4 (极限情形)。 当 $s \to t$时，函数 h_t 收敛至：

$$h_t(\mathbf{x}, \mathbf{z}_t, \tilde{\mathbf{x}}_0) \approx \frac{p_{\mathbf{x}, \mathbf{z}_t} - p_{\tilde{\mathbf{x}}_0, \mathbf{z}_t}}{f_t(\tilde{\mathbf{x}}_0, \mathbf{z}_t)} - \frac{1}{\alpha_t} \log \frac{f_t(\mathbf{x}, \mathbf{z}_t)}{f_t(\tilde{\mathbf{x}}_0, \mathbf{z}_t)} + \log \frac{f_t(\mathbf{x}, \mathbf{x})}{f_t(\tilde{\mathbf{x}}_0, \mathbf{x})} + \frac{k \beta_t}{N \alpha_t} \sum_{\mathbf{e} \in \mathcal{V}} \log \frac{f_t(\mathbf{x}, \mathbf{e})}{f_t(\tilde{\mathbf{x}}_0, \mathbf{e})}. \tag{54}$$

证明。 注意，当 $s \to t$, $\beta_{t|s}\alpha_s \to 0$ 时，这是一个 $0/0$ 商。应用洛必达法则可得。（分子和分母均

对 $\alpha_{t|s}$ 求导），我们有

$$\lim_{s\to t}\frac{1}{\beta_{t|s}\alpha_s}\log\frac{f_s(\mathbf{x},\ \mathbf{z_t})f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})f_s(\tilde{}\mathbf{x}_0,\mathbf{z_t})} \tag{55}$$

$$=\frac{-1}{\alpha_s}\frac{\partial}{\partial\alpha_{t|s}}\log\frac{f_s(\mathbf{x},\mathbf{z_t})f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})f_s(\tilde{}\mathbf{x}_0,\mathbf{z_t})} \tag{56}$$

$$=\frac{-1}{\alpha_s}\left(\frac{\partial\log f_s(\mathbf{x},\mathbf{z_t})f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}{\partial\alpha_{t|s}}-\frac{\partial\log f_t(\mathbf{x},\mathbf{z_t})f_s(\tilde{\mathbf{x}}_0,\mathbf{z_t})}{\partial\alpha_{t|s}}\right) \tag{57}$$

$$=\frac{-1}{\alpha_s}\left(\frac{\partial\log f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}{\partial\alpha_{t|s}}-\frac{\partial\log f_t(\mathbf{x},\mathbf{z_t})}{\partial\alpha_{t|s}}\right) \tag{58}$$

$$=\frac{-1}{\alpha_s}\left(\frac{1}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}\frac{\partial(\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}+(1-\alpha_t)r(\mathbf{z_t}))}{\partial\alpha_{t|s}}-\frac{1}{f_t(\mathbf{x},\mathbf{z_t})}\frac{\partial f_t(\mathbf{x},\mathbf{z_t})}{\partial\alpha_{t|s}}\right) \tag{59}$$

$$=\frac{-1}{\alpha_s}\left(\frac{1}{f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})}\left(\alpha_s p_{\tilde{\mathbf{x}}_0,\ \mathbf{z_t}}-\alpha_s r(\mathbf{z_t})\right)-\frac{1}{f_t(\mathbf{x},\mathbf{z_t})}\frac{\partial f_t(\mathbf{x},\mathbf{z_t})}{\partial\alpha_{t|s}}\right) \tag{60}$$

$$=-\frac{p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}-r(\mathbf{z_t})}{f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})}+\frac{p_{\mathbf{x},\mathbf{z_t}}-r(\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})} \tag{61}$$

$$=\frac{(\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}+\beta_t r(\mathbf{z_t}))(p_{\mathbf{x}\ ,\mathbf{z_t}}-r(\mathbf{z_t}))\ -(\alpha_t p_{\mathbf{x}\ ,\mathbf{z_t}}+\beta_t r(\mathbf{z_t}))(p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}-r(\mathbf{z_t}))}{f_t(\mathbf{x},\mathbf{z_t})f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})} \tag{62}$$

$$=\frac{-\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}r(\mathbf{z_t})+\beta_t p_{\mathbf{x},\mathbf{z_t}}r(\mathbf{z_t})+\alpha_t p_{\mathbf{x},\mathbf{z_t}}r(\mathbf{z_t})-\beta_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}r(\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})} \tag{63}$$

$$=\frac{r(\mathbf{z_t})(p_{\mathbf{x}\ ,\mathbf{z_t}}-p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}})}{f_t(\mathbf{x},\mathbf{z_t})f_t(\tilde{}\mathbf{x}_0,\mathbf{z_t})} \tag{64}$$

so

$$\lim_{s\to t}\left(\frac{f_s(\mathbf{x},\mathbf{z_t})}{r(\mathbf{z_t})}\frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s}\log\frac{f_s(\mathbf{x},\mathbf{z_t})f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})f_s(\tilde{\mathbf{x}}_0,\mathbf{z_t})}\right)=\frac{p_{\mathbf{x},\mathbf{z_t}}-p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} \tag{65}$$

□

C. 与 MDLM 和 UDLM 的关系

引理 5. 通过设置 $k=0$，*XDLM* 可以简化为 *MDLM*，其后验概率为：

$$p_\theta(\mathbf{z_s}=\mathbf{e}|\mathbf{z_t})=\delta_{\mathbf{z_t},\mathbf{m}}\delta_{\mathbf{e},\mathbf{m}}\frac{\beta_s}{\beta_t}+\delta_{\mathbf{z_t},\mathbf{m}}\delta_{\mathbf{e}\neq\mathbf{m}}\frac{\beta_{t|s}\alpha_s p_{\theta,\mathbf{e}}}{\beta_t}+\delta_{\mathbf{z_t}\neq\mathbf{m}}\delta_{\mathbf{e},\mathbf{z_t}} \tag{66}$$

$$=\begin{cases}\frac{\beta_s}{\beta_t}\delta_{\mathbf{m},\mathbf{z_t}}, & \mathbf{e}=\mathbf{m}\\ \frac{(\alpha_s-\alpha_t)}{1-\alpha_t}p_{\theta,\mathbf{e}}, & \mathbf{z_t}=\mathbf{m},\mathbf{e}\neq\mathbf{m}\\ \delta_{\mathbf{e},\mathbf{z_t}}, & \mathbf{z_t}\neq\mathbf{m},\mathbf{e}\neq\mathbf{m}\end{cases} \tag{67}$$

以及 *KL* 散度

$$KL=-\frac{\beta_{t|s}\alpha_s}{\beta_t}\log p_{\theta,\mathbf{x}} \tag{68}$$

证明。由于 $k=0$、$\mu=1$，我们有

$$r(\mathbf{e})=\delta_{\mathbf{e},\mathbf{m}} \tag{69}$$

$$p_{\tilde{\mathbf{x}}_0,\mathbf{e}}=\delta_{\mathbf{e}\neq\mathbf{m}}p_{\tilde{\mathbf{x}}_0,\mathbf{e}} \tag{70}$$

因此

$$q(\mathbf{z_s}=\mathbf{e}|\mathbf{z_t},\ \tilde{\mathbf{x}}_0) = \frac{f_s(\tilde{\mathbf{x}}_0,\mathbf{e})f_{t|s}(\mathbf{e},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} \tag{71}$$

$$= \frac{(\alpha_s p_{\tilde{\mathbf{x}}_0,\mathbf{e}}+\beta_s r(\mathbf{e}))(\alpha_{t|s}p_{\mathbf{e},\mathbf{z_t}}+\beta_{t|s}r(\mathbf{z_t}))}{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z}_\mathbf{t}}+\beta_t r(\mathbf{z_t})} \tag{72}$$

$$= \frac{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{e}}\delta_{\mathbf{e},\mathbf{z_t}}+\alpha_s\beta_{t|s}p_{\tilde{\mathbf{x}}_0,\mathbf{e}}\delta_{\mathbf{z_t},\mathbf{m}}+\alpha_{t|s}\beta_s\delta_{\mathbf{e},\mathbf{z_t}}\delta_{\mathbf{e},\mathbf{m}}+\beta_s\beta_{t|s}\delta_{\mathbf{e},\mathbf{m}}\delta_{\mathbf{z_t},\mathbf{m}}}{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}+\beta_t\delta_{\mathbf{z_t},\mathbf{m}}} \tag{73}$$

$$= \delta_{\mathbf{z_t},\mathbf{m}}\frac{\alpha_s\beta_{t|s}p_{\tilde{\mathbf{x}}_0,\mathbf{e}}+\alpha_{t|s}\beta_s\delta_{\mathbf{e},\mathbf{z_t}}+\beta_s\beta_{t|s}\delta_{\mathbf{e},\mathbf{z_t}}}{\beta_t} \tag{74}$$

$$+\delta_{\mathbf{z_t}\neq\mathbf{m}}\ \frac{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{e}}\delta_{\mathbf{e},\mathbf{z_t}}}{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z}_\mathbf{t}}} \tag{75}$$

$$= \delta_{\mathbf{z_t},\mathbf{m}}\left(\frac{\alpha_s\beta_{t|s}}{\beta_t}p_{\tilde{\mathbf{x}}_0,\mathbf{e}}+\frac{\beta_s}{\beta_t}\delta_{\mathbf{e},\mathbf{z_t}}\right)+\delta_{\mathbf{z_t}\neq\mathbf{m}}\delta_{\mathbf{e},\mathbf{z_t}}\frac{p_{\tilde{\mathbf{x}}_0,\mathbf{e}}}{p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}} \tag{76}$$

$$= \delta_{\mathbf{z_t},\mathbf{m}}\frac{\alpha_s\beta_{t|s}}{\beta_t}p_{\tilde{\mathbf{x}}_0,\mathbf{e}}+\frac{\beta_s}{\beta_t}\delta_{\mathbf{z_t},\mathbf{m}}\delta_{\mathbf{e},\mathbf{z_t}}+\delta_{\mathbf{z_t}\neq\mathbf{m}}\delta_{\mathbf{e},\mathbf{z_t}} \tag{77}$$

对于 kl，我们有

$$h_t(\mathbf{x},\mathbf{z_t}) = \frac{f_s(\mathbf{x},\mathbf{z_t})}{r(\mathbf{z_t})}\frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s}\log\frac{f_s(\mathbf{x},\mathbf{z_t})f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})f_s(\tilde{\mathbf{x}}_0,\mathbf{z_t})} \tag{78}$$

$$-\frac{1}{\alpha_s}\log\frac{f_t(\mathbf{x},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}+\log\frac{f_s(\mathbf{x},\mathbf{x})}{f_s(\tilde{\mathbf{x}}_0,\mathbf{x})}+\frac{k\beta_s}{N\alpha_s}\sum_{\mathbf{e}}\log\frac{f_s(\mathbf{x},\mathbf{e})}{f_s(\tilde{\mathbf{x}}_0,\mathbf{e})} \tag{79}$$

$$= \frac{f_s(\mathbf{x},\mathbf{z_t})}{r(\mathbf{z_t})}\frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s}\log\frac{(\alpha_s p_{\mathbf{x},\mathbf{z_t}}+\beta_s\delta_{\mathbf{z_t},\mathbf{m}})(\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}+\beta_t\delta_{\mathbf{z_t},\mathbf{m}})}{(\alpha_t p_{\mathbf{x},\mathbf{z_t}}+\beta_t\delta_{\mathbf{z_t},\mathbf{m}})(\alpha_s p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}+\beta_s\delta_{\mathbf{z_t},\mathbf{m}})} \tag{80}$$

$$-\frac{1}{\alpha_s}\log\frac{\alpha_t p_{\mathbf{x},\mathbf{z_t}}+\beta_t\delta_{\mathbf{z_t},\mathbf{m}}}{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z}_\mathbf{t}}+\beta_t\delta_{\mathbf{z_t},\mathbf{m}}}-\text{对数}p_{\tilde{\mathbf{x}}_0,\ \mathbf{x}} \tag{81}$$

$$= \frac{f_s(\mathbf{x},\mathbf{z_t})}{r(\mathbf{z_t})}\frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s}\log\left(\delta_{\mathbf{z_t},\mathbf{m}}\frac{\beta_s\delta_{\mathbf{z_t},\mathbf{m}}\beta_t\delta_{\mathbf{z_t},\mathbf{m}}}{\beta_t\delta_{\mathbf{z_t},\mathbf{m}}\beta_s\delta_{\mathbf{z_t},\mathbf{m}}}+\delta_{\mathbf{z_t}\neq\mathbf{m}}\frac{\alpha_s p_{\mathbf{x},\mathbf{z_t}}\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}{\alpha_t p_{\mathbf{x},\mathbf{z_t}}\alpha_s p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}\right) \tag{82}$$

$$-\frac{1}{\alpha_s}\log\left(\delta_{\mathbf{z_t},\mathbf{m}}\frac{\beta_t\delta_{\mathbf{z_t},\mathbf{m}}}{\beta_t\delta_{\mathbf{z_t},\mathbf{m}}}+\delta_{\mathbf{z_t}\neq\mathbf{m}}\frac{\alpha_t p_{\mathbf{x},\mathbf{z_t}}}{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}\right)-\log p_{\tilde{\mathbf{x}}_0,\mathbf{x}} \tag{83}$$

$$= \frac{f_s(\mathbf{x},\mathbf{z_t})}{r(\mathbf{z_t})}\frac{\alpha_{t|s}}{\beta_{t|s}\alpha_s}\log(1)-\frac{1}{\alpha_s}\delta_{\mathbf{z_t}\neq\mathbf{m}}\log\frac{p_{\mathbf{x},\mathbf{z_t}}}{p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}-\log p_{\tilde{\mathbf{x}}_0,\mathbf{x}} \tag{84}$$

so

$$D_{KL} = \frac{\beta_{t|s}\alpha_s r(\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})}h_t(\mathbf{x},\mathbf{z_t}) \tag{85}$$

$$= -\frac{\beta_{t|s}\delta_{\mathbf{z_t},\mathbf{m}}}{f_t(\mathbf{x},\mathbf{z_t})}\delta_{\mathbf{z_t}\neq\mathbf{m}}\log\frac{p_{\mathbf{x},\mathbf{z_t}}}{p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}-\frac{\beta_{t|s}\alpha_s\delta_{\mathbf{z_t},\mathbf{m}}}{\alpha_t\delta_{\mathbf{x},\mathbf{z_t}}+\beta_t\delta_{\mathbf{z_t},\mathbf{m}}}\log p_{\tilde{\mathbf{x}}_0,\mathbf{x}} \tag{86}$$

$$= -\delta_{\mathbf{z_t},\mathbf{m}}\frac{\beta_{t|s}\alpha_s}{\beta_t}\log p_{\tilde{\mathbf{x}}_0,\mathbf{x}} \tag{87}$$

□

引理6。 设置 $k=1$ 后，*XDLM* 可以化简为 *UDLM*，其后验概率为：

$$p_\theta(\mathbf{z_s}=\mathbf{e}|\mathbf{z_t}) = \frac{\delta_{\mathbf{e},\mathbf{z_t}}(N\alpha_t p_{\theta,\mathbf{e}}+\beta_s\alpha_{t|s})+(\alpha_s-\alpha_t)p_{\theta,\mathbf{e}}+\beta_s\beta_{t|s}/N}{N\alpha_t p_{\theta,\mathbf{z_t}}+\beta_t} \tag{88}$$

以及 KL散度

$$KL = \frac{-\beta_{t|s}\alpha_s}{N\alpha_t}\left(\frac{1}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \frac{1}{f_t(\mathbf{x},\mathbf{z_t})} - \sum_{\mathbf{e}\neq\mathbf{z_t}}\frac{f_t(\mathbf{x},\mathbf{e})}{f_t(\mathbf{x},\mathbf{z_t})}\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})f_t(\mathbf{x},\mathbf{z_t})}\right) \tag{89}$$

证明。 由于 $k=1$、 $\mu=0$，我们有

$$\forall\mathbf{e},\ r(\mathbf{e}) = r = \frac{1}{N} \tag{90}$$

因此

$$q(\mathbf{z_s}=\mathbf{e}|\mathbf{z_t},\tilde{\mathbf{x}}_0) = \frac{f_s(\tilde{\mathbf{x}}_0,\mathbf{e})f_{t|s}(\mathbf{e},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} \tag{91}$$

$$= \frac{(\alpha_s p_{\tilde{\mathbf{x}}_0,\mathbf{e}} + \beta_s/N)(\alpha_{t|s}p_{\mathbf{e},\mathbf{z_t}} + \beta_{t|s}/N)}{\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}} + \beta_t/N} \tag{92}$$

$$= \frac{N\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{e}}\delta_{\mathbf{e},\mathbf{z_t}} + \alpha_s\beta_{t|s}p_{\tilde{\mathbf{x}}_0,\mathbf{e}} + \alpha_{t|s}\beta_s\delta_{\mathbf{e},\mathbf{z_t}} + \beta_s\beta_{t|s}/N}{N\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}} + \beta_t} \tag{93}$$

$$= \frac{\delta_{\mathbf{e},\mathbf{z_t}}(N\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{e}} + \alpha_{t|s}\beta_s) + (\alpha_s-\alpha_t)p_{\tilde{\mathbf{x}}_0,\mathbf{e}} + \beta_s\beta_{t|s}/N}{N\alpha_t p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}} + \beta_t} \tag{94}$$

给定 $s\to t$，对于 KL，我们有

$$h_t(\mathbf{x},\mathbf{z_t}) \tag{95}$$

$$= \frac{p_{\mathbf{x},\mathbf{z_t}} - p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \frac{1}{\alpha_t}\log\frac{f_t(\mathbf{x},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} + \log\frac{f_t(\mathbf{x},\mathbf{x})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{x})} + \frac{k\beta_t}{N\alpha_t}\sum_{\mathbf{e}}\log\frac{f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})} \tag{96}$$

$$= f_t(\mathbf{x},\mathbf{z_t})\frac{p_{\mathbf{x},\mathbf{z_t}} - p_{\tilde{\mathbf{x}}_0,\mathbf{z_t}}}{f_t(\mathbf{x},\mathbf{z_t})f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \frac{1}{\alpha_t}\log\frac{f_t(\mathbf{x},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} + \log\frac{f_t(\mathbf{x},\mathbf{x})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{x})} \tag{97}$$

$$+ \frac{\beta_t}{N\alpha_t}\sum_{\mathbf{e}}\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})f_t(\mathbf{x},\mathbf{z_t})} - \frac{\beta_t}{N\alpha_t}\sum_{\mathbf{e}}\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})} \tag{98}$$

$$= \frac{f_t(\mathbf{x},\mathbf{z_t})}{\alpha_t}\left(\frac{1}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \frac{1}{f_t(\mathbf{x},\mathbf{z_t})}\right) - \log\frac{f_t(\mathbf{x},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} + \log\frac{f_t(\mathbf{x},\mathbf{x})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{x})} \tag{99}$$

$$+ \frac{\beta_t}{N\alpha_t}\sum_{\mathbf{e}}\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})f_t(\mathbf{x},\mathbf{z_t})} \tag{100}$$

$$= \frac{f_t(\mathbf{x},\mathbf{z_t})}{\alpha_t}\left(\frac{1}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \frac{1}{f_t(\mathbf{x},\mathbf{z_t})}\right) - \log\frac{f_t(\mathbf{x},\mathbf{z_t})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} + \log\frac{f_t(\mathbf{x},\mathbf{x})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{x})} \tag{101}$$

$$+ \frac{1}{\alpha_t}\sum_{\mathbf{e}} f_t(\mathbf{x},\mathbf{e})\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})f_t(\mathbf{x},\mathbf{z_t})} - \log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{x})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{x})f_t(\mathbf{x},\mathbf{z_t})} \tag{102}$$

$$= \frac{f_t(\mathbf{x},\mathbf{z_t})}{\alpha_t}\left(\frac{1}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \frac{1}{f_t(\mathbf{x},\mathbf{z_t})}\right) + \frac{1}{\alpha_t}\sum_{\mathbf{e}} f_t(\mathbf{x},\mathbf{e})\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})f_t(\mathbf{x},\mathbf{z_t})} \tag{103}$$

so

$$D_{KL} = \frac{\beta_{t|s}\alpha_s r(\mathbf{z_t})}{f_t(\mathbf{x},\mathbf{z_t})}h_t(\mathbf{x},\mathbf{z_t}) \tag{104}$$

$$= \frac{-\beta_{t|s}\alpha_s}{N\alpha_t}\left(\frac{1}{f_t(\mathbf{x},\mathbf{z_t})} - \frac{1}{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})} - \sum_{\mathbf{e}}\frac{f_t(\mathbf{x},\mathbf{e})}{f_t(\mathbf{x},\mathbf{z_t})}\log\frac{f_t(\tilde{\mathbf{x}}_0,\mathbf{z_t})f_t(\mathbf{x},\mathbf{e})}{f_t(\tilde{\mathbf{x}}_0,\mathbf{e})f_t(\mathbf{x},\mathbf{z_t})}\right) \tag{105}$$

□

D. LM1B 采样情形

我们研究推理计算预算（以采样步数 T 定义）与生成文本的感知质量之间的关系。表 5 展示了从基于 LM1B 训练的 XDLM 模型中随机选取的未经筛选的采样。每个采样代表一个独立扩散过程在特定总步数 $T \in \{4, 8, 16, 32, 64, 128\}$ 下的最终输出。

在采样预算的下限（$T = 4$）下，模型无法收敛到数据流形。输出表现为彼此割裂的词汇检索（例如：“广告 - 标签尖头”），其中各个词元本身有效，但缺乏句法绑定。将预算增加至 $T = 8$ 后，开始出现表层句法结构（例如：“*Curt Schilling* 是一个新颖性十足的人”），但语义内容仍不合逻辑，这表明扩散过程会先解决局部语法依存关系，再处理更高阶的语义意义。

随着采样步数增加至中间区间（$T = 16$ 和 $T = 32$），我们观察到文本从仅仅语法正确逐渐过渡到叙事合理。在 $T = 16$ 时，文本呈现出有效的句子结构，但存在同义反复和主题漂移的问题（例如：“中东在中东”）。到了 $T = 32$，模型能够生成连贯的分句，且主谓宾关系清晰，尽管具体实体往往仍是合成或幻觉出来的（例如：“*Poundhead* 先生”）。这表明，虽然 32 步足以让模型学会语言的结构规则，但去噪过程尚未使输出与训练语料库的具体事实分布完全对齐。

在 $T = 64$ 处观察到样本质量出现明显的相变，这似乎代表了高保真生成的饱和点。在此阶段，模型能够生成语义稳健、与自然新闻数据难以区分的文本，并成功处理复杂的实体列表（例如“刚果、贝宁、科特迪瓦”）和特定领域术语（例如“美国国家运输安全委员会”）。将过程延长至 $T = 128$ 后，长程一致性和定量数据的精确度略有提升，但与 $T = 64$ 相比，感知上的提升远小于从 $T = 32$ 开始所观察到的跃升。这一轨迹表明，XDLM 需要至少约 64 步才能解析细粒度的语义细节；但进一步增加计算成本，对生成质量的提升会逐渐减小。

表 *5.* **XDLM 在 LM1B 上训练生成的未筛选样本**。每个样本代表一个独立扩散过程的输出，该过程受特定总步数 T 的约束。序列长度固定为 128 个词元。

步数（T）	生成文本样本
$T = 4$	[CLS]名词 [CLS] 这些前候选人的广告标签是 九月。 [CLS] 新的和有抱负的竞争者吹嘘一份礼物，令人困惑 不是候选人，不仅仅是为了15 - u运行债券。 [CLS] 乘客在 董事会波多黎各喷气式飞机喷气飞往灰烬oftwa，夏威夷，星期三那个 下午。 [CLS] 由于融合从马力中进行并且 献给奥克特和公羊，以及通用汽车，租赁收益／三分之一是陷阱。 [CLS] 时代广场可能会经历复辟 [CLS] 而“旧安拉”是一个 向“那一个，那一个”致敬 在白雪皑皑的伊斯兰信仰中，“上帝”。 [CLS] 不幸的是 Christina 看见 ， [CLS]
$T = 8$	[CLS] 太胖了，无法过去。 [CLS] Curt Schilling是一个富有新颖性的人。 [CLS] Middle先生曾有， 15年来一直没有开始做这门生意， 想知道董事会中是否有人曾经想从卡夫那里得到某样东西。 [CLS] 这位新近频繁受任的所有者于星期三提议，政府 应从财政部收取“惩罚性税收附加款”，并坚持要求 印第安纳州和密歇根州遵守通用汽车的合同条款。 [CLS] 卡明斯女士 : 有时，人们会决定接受复辟。 [CLS] 而对于 如今，巴基斯坦议会已甘愿接受民主化的动荡 投票告别扎尔达里，它认为自己是 [CLS]

步数（T）	生成文本样本（续）
$T = 16$	[CLS] ”未能通过广告限制。 [CLS] 该网络，其中是中东唯一已知的新事物，预期用于大约10到15年。 [CLS] 我相信董事会的人曾非常高兴。 我如何能够在自己内心解释它。 [CLS] Studio 86几乎大量当代电视转播的——汽车展——招牌——汽车。 前期 - 通用汽车的生物燃料尝起来像三分之二的标志性美国风味: 三分之一的福特使用的时间，约 4 年后，原版又回到了建筑中它历经多年。 [CLS] 相比之下，扎尔达里认为华盛顿 [CLS]
$T = 32$	[CLS] 我，以便结清账目。 [CLS] 这种中立感很可能迫使争端多次延期，而争端已于星期三解决。 [CLS] 15 告诉《每日电讯报》，他一直认为董事会已经说服他该坐火车了。 [CLS] Poundhead先生想要市议会创建一系列新的游客区。 [CLS] Marfor，据《世界报》报道， "是阿富汗人偷走或拿走并带走的" - - 塔利班经常用来从巴基斯坦塔利班手中购买的资金，而非巴基斯坦的。 [CLS] 尽管扎尔达里说，他在西方有一些朋友看到 k [CLS]
$T = 64$	[CLS] 我，XV230号机组人员，这架飞机载有刚果、贝宁、科特迪瓦佛得角、曼彻斯特、俄罗斯人以及罗安达的海岸警卫队。 [CLS] 美国国家运输安全委员会曾将电气生命将支撑背心运送到夏威夷上空的安全地点。 [CLS] 《每日先驱报》称分析师预计收益为3.94亿美元，即每股43美分，由于迷你市场的增长好于预期。 [CLS] 卡迪将军写道，这些管控措施可用于防止报复性袭击，但并非为了阻止那些被怀疑扰乱网络的人发起抗议岛屿上的，提出了这一理论。 [分类标记] [分类标记]
$T = 128$	[CLS] 让我代表他们所代表的古巴人民。 [CLS] 该领导层仍在接管公司，但问题已经尚未得到解决。 [CLS] 对于 EEY Investments 的 Tony Fernanda Santos 而言董事会认为，这造成了 3.28 亿英镑的税务亏损，并带来 8,600 万美元的收益，以及预计到 2011 年下一年的 3,700 万美元收益。 [CLS] 接下来是 “最佳作品揭示， “第一部迷你剧。 [CLS] cusuoga，这种常用于描述报复性袭击的民族主义方法，在大多数反对猛虎组织存在的抗议中并未出现该岛屿的僧伽罗地区。 [分类标记] [分类标记]

为探究 XDLM 的内部生成机制，我们将单条生成轨迹可视化，其总预算为 $T = 32$ 步，序列长度固定为 128 个词元。表 6 详细展示了序列在以下步数的演变：$t \in \{0, 1, 8, 16, 24, 32\}$。我们采用颜色编码方案，以区分混合噪声过程中固有的具体转移类型。**绿色**表示从吸收态（[MASK]）转移到词元（掩码扩散语言模型的典型特征），**蓝色**表示词元到词元细化（均匀扩散语言模型的典型特征），而**红色**表示重掩码操作（词元退回到[MASK]）。这种红色转移是 XDLM 特有的；其前向过程由吸收噪声和均匀噪声混合而成，因此模型能够通过随机拒绝先前生成的词元来回溯。

过程在 $t = 0$处开始，混合了掩码以及源自均匀噪声的随机词元，例如“棺材”、“湿滑”和“小号”。在向 $t = 1$ 过渡时，立即观察到一个关键动态。模型主动拒绝了这些随机初始化中的大多数。我们观察到大量**红色**标记，其中像“棺材”和“小号”这样的词元被还原为 [MASK]。这表明 XDLM 并不严格受其来自均匀噪声的随机初始化所束缚。与标准迭代细化模型可能难以摆脱由不良随机种子引起的局部极小值不同，XDLM 具备识别低概率词元并通过重新掩码“推倒重来”的能力。与此同时，有效的锚点词元开始出现（**绿色**），例如“先令”和“受阻”，从而确立了初始语义方向。

随着生成在中间步数中推进（$t = 1 \to t = 16$），模型在生成与纠错之间展开动态交互。三种转换类型全部出现，表明这是一个非单调搜索过程。模型会生成新候选（**绿色**），并优化已有候选（**蓝色**），但关键在于，它会继续利用重新掩码（**红色**）来剪除不一致分支。例如，在早期步数中暂定生成的词元，会在它们

未能与不断演变的全局上下文对齐时。

在最终阶段（$t = 24 \to t = 32$），其动态呈现出鲜明的变化。**红色**重新掩码转换完全消失，表明结构收敛。该过程专注于填补剩余空缺（**绿色**）并进行细粒度词汇替换（**蓝色**）。此阶段会出现显著的修正，例如第 24 步的词元“*life infancy*”在第 32 步被修正为“*was time*”，以及“*from pound*”被修正为“*from pound-land*”。这一轨迹证实，XDLM 成功地从以主动删除和重新生成为特征的高温探索状态退火至低温细化状态，从而确保最终输出兼具整体连贯性和局部精确性。

表 6. **生成序列的逐步演变**（$T = 32$）。文本颜色表示混合噪声过程中固有的转移动力学：**绿色**表示由掩码生成的新词元；**蓝色**表示词汇细化；而**红色**突出显示重掩码操作：此前生成的词元被舍弃，并恢复为 [掩码]。

步数（T）	生成文本样本
$T = 0$	[掩码] [掩码] 踉跄 [掩码] [掩码] 棺材 [掩码] [掩码] [掩码] [掩码] 也 [掩码] [掩码] [掩码] [掩码] [掩码] [掩码] [掩码] [掩码] [掩码] [掩码] [遮罩] [遮罩] [遮罩] [遮罩] [遮罩] [遮罩] [遮罩] [遮罩] , [掩码] [掩码] [MASK] [MASK] [MASK] [MASK] [MASK] ##agawa [MASK] [MASK] [MASK] [MASK] 滑溜的 [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] Harrington [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] 小号超越 [MASK] 拒绝 [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] ##was 老兵 一月 [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] begs [MASK] [MASK]
$T = 1$	[MASK] [MASK] 步履蹒跚 [MASK] [MASK] **[MASK]** [MASK] [MASK] [MASK] [MASK] **[MASK]** **##fo** [MASK] **是（初始分数）** [MASK] [MASK] **先令** [MASK] [MASK] [MASK] [MASK] [掩码] [掩码] [掩码] [掩码] [掩码] [掩码] **1879** [掩码] **[掩码]** [掩码] [掩码] [MASK] **##anda** [MASK] **受阻** [MASK] ##agawa [MASK] [MASK] [MASK] [MASK] **[MASK]** [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] **[MASK]** [MASK] **MASK** [MASK] [MASK] [MASK] **一台** [[掩码]] [[掩码]] **##itive** **[[掩码]]** **[[掩码]]** [[掩码]] [[掩码]] [[掩码]] **mar** [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] **分发** [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] **[[掩码]]** [[掩码]] [[掩码]] [[掩码]]**三级** [[掩码]] [[掩码]] [[掩码]] [[掩码]] **[[掩码]]** **[[掩码]]** **[[掩码]]** [[掩码]] [[掩码]] [[掩码]] **大步走过** [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] [[掩码]] **无论什么** [[掩码]] **##rda** [[掩码]] [掩码] [掩码] **[掩码]** [掩码] [掩码]
$T = 8$	[MASK] [MASK] **[MASK]** **清除** [MASK] [MASK] [MASK] **[CLS]** [MASK] [MASK] **的** [MASK] 是 [MASK] **刚果** **[MASK]** [MASK] [MASK] [MASK] [MASK] [MASK] **##rians** **圣诞老人** **，** [MASK] [MASK] **[MASK]** [MASK] [MASK] [MASK] **15** [MASK] ##anda [MASK] **[MASK]** [MASK] **[MASK]** [MASK] [MASK] **董事会有** [MASK] **##ffi** [MASK] [MASK] [MASK] **##illo** [MASK] **该** [MASK] [MASK] **[CLS]** [MASK] **夏威夷** [MASK] [MASK] **该** **[MASK]** **86** **到** [MASK] 一 [MASK] **的** **过程** [MASK] [MASK] [MASK] [MASK] **[CLS]** mar [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] **，** [遮罩] [遮罩] [遮罩] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]**巴西** [MASK] [MASK] [MASK] [MASK] [MASK] **该** [MASK] [MASK] [MASK] **使用** **[MASK]** [MASK] [MASK] **巴基斯坦人** [MASK] [MASK] [MASK] [MASK] **’** **s** [MASK] **[MASK]** **##west** [MASK] [MASK] [MASK] [MASK] **在** [MASK] **[MASK]** **za** ##rda [MASK] [MASK] [MASK] **看见** [掩码] [掩码]

步数（T）	生成文本样本（续）
$T = 16$	[MASK] [MASK] [MASK] 清除 [MASK] [MASK] [MASK] [CLS] [MASK] [MASK] 的 [MASK] 是 [MASK] 刚果 [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] **the** **[MASK]** ，**其中** [MASK] **已解决淫秽内容**。 **[CLS]** 15 [MASK] ##anda [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] 董事会 had **convinced** ##ffi [MASK] **life** [MASK] **[MASK]** **take** the [MASK] [MASK] [CLS] [MASK] 夏威夷 [MASK] [MASK] the **城市** 86to [MASK] a [MASK] of **新** [MASK] [MASK] **地区** [MASK] [CLS] mar [MASK] **副** [MASK] [MASK] **le** [MASK] ， [MASK] [MASK] [MASK] **阿富汗人** [MASK] **窃取** [MASK] [MASK] [MASK] **[MASK]** **出** [MASK] **–** [MASK] [MASK] 该 **塔利班经常使用来** [MASK] **从巴基斯坦人** [MASK] [MASK] **不是巴基斯坦** ’ s [MASK] [MASK] **尽管** [MASK] **其** [MASK] 在 [MASK] [MASK] za ##rda **##ri说** [MASK] 看到 [MASK] [MASK]
$T = 24$	[MASK] [MASK] [MASK] 清除 [MASK] [MASK] 。 [CLS] **该** [MASK] 的 **中立**是 [MASK] 刚果**迫使一些延期**该 [MASK] ，该 [掩码] 解决了 **[掩码]** 。 [分类标记] 15 [掩码]**《每日电讯报》所** [掩码] [掩码] **相信**董事会已经说服了**他** [掩码] 生命**婴儿期** [掩码] 乘坐**火车** [掩码] [分类标记] **从庞德** [掩码] **，** 这座城市**计划**要 [掩码] 一个 [掩码] 新 [掩码] [掩码] 地区 [掩码] [分类标记] mar [掩码] **，** **据** [掩码] 世界**报**， **“是**阿富汗人**##s**偷窃**或** [掩码]**以及脚**离开”**–** **–** [掩码] 塔利班经常用来 [掩码] 来自巴基斯坦**塔利班** [掩码] 不是巴基斯坦的 [掩码] [掩码] 尽管 [掩码] 其**朋友**中的**那些** [掩码] 扎尔达里说**他**认为 [掩码] [掩码]
$T = 32$	**[分类标记] 我，** 以澄清**旗帜**。 [分类标记] 这种**类型**的中立**可能** **以**迫使某些推迟**争端**，该争端**已被**解决 **星期三**。 [CLS] 15 **告诉**《每日电讯报》，**他一直**相信该 董事会已说服他，**是时候去**乘火车。 [CLS] 来自庞德 **##land** ，这座城市计划**建造一系列**新的**零售住宅**地区 · [CLS] mar**##for** ，据《世界报》， “是阿富汗人用来偷窃的 或者 **拿走** 并带走“—— **钱财** 塔利班经常用来 **购买** 来自 巴基斯坦塔利班， 而不是巴基斯坦的 。 **[CLS]** 尽管 **其中一些** 朋友 在 **西方**扎尔达里说他看到 **k** **[CLS]**

E. ImageNet采样案例

我们在 ImageNet-1K 上对 XDLM 与基线模型 MDLM、GIDD 和 UDLM 进行了定性比较。结果分两种设置进行分析：无引导生成（图 5）以及采用轻微无分类器引导的生成（图 6）。视觉证据凸显了 XDLM 的独特能力：它能够兼取吸收式策略和均匀噪声策略的优势，同时减轻各自的缺点。

无引导设置是对结构先验项的压力测试。在这种情况下，MDLM 和 GIDD 的局限性最为明显。这些模型难以形成连贯几何结构，产生杂乱的纹理团块（例如火烈鸟和瓢虫所在列），而非清晰可辨的物体。相比之下，仅依赖均匀噪声的 UDLM 能够成功捕捉整体形状，却无法生成高频细节，因此图像模糊且失焦。

XDLM 通过平衡这两个极端展现出独特优势。通过融合 MDLM 的结构刚性与 UDLM 的细化灵活性，XDLM 在不牺牲清晰度的同时实现了结构连贯性。例如，在火箭发射案例中，XDLM 是唯一一个在无引导条件下呈现出清晰垂直机身和鲜明烟柱的模型。

如图 6 所示，引入 2.0 的引导尺度可提升整体样本质量，但各模型的典型表现仍然存在。MDLM 仍受细节细化不足的困扰，常常生成过度饱和的纹理，语义结构也较差。UDLM 依然过于平滑；图像虽然干净，却缺乏照片级真实感所需的锐利感。

XDLM 最有效地利用引导来增强细节。火烈鸟群的例子充分体现了这一点。XDLM 能清晰呈现一只只羽毛锐利、彼此分明的鸟，而 MDLM 将它们融合成一团，UDLM 则模糊了边界。同样，在食物类别（草莓和披萨）中，XDLM 生成了逼真的表面纹理和光照反射，而这些细节在 UDLM 的采样结果中缺失，在 MDLM 的采样结果中则发生了扭曲。这证实了 XDLM 的混合噪声

该机制为高保真生成提供了更优的基础。



图 *5.* **无 CFG 条件下 ImageNet-1K 上类别条件生成的定性比较。** 在没有引导的情况下，基线模型表现明显不佳：MDLM 和 GIDD 产生了杂乱的伪影，结构连贯性较差，而 UDLM 生成的图像虽可辨认，但过于平滑。XDLM 有效平衡了 MDLM 和 UDLM 的特性，即使没有引导，也能生成最连贯且语义正确的样本（例如清晰的火箭结构和披萨配料）。



图 *6.* **CFG 尺度为 2.0 时 ImageNet-1K 上类别条件生成的定性比较。** 加入引导后，所有模型的表现均有所提升，但质量差异依然存在。MDLM 和 GIDD 仍然表现出纹理失真和“烧焦”伪影。UDLM 生成的结果干净但模糊，缺乏细粒度纹理。XDLM 展现出更高的保真度，将清晰的高频细节（如草莓和火山上的细节）与准确的全局结构相结合。

我们进一步分析 XDLM（$k = 0.1$）在 ImageNet-1K 上以 CFG 为 2.0 进行 8 步生成的采样动态，并将其与 MDLM、GIDD 和 UDLM 进行比较。视觉演变展现了各模型处理噪声和细化特征时的不同表现。

XDLM 的一项关键优势是其强大的细化能力，这得益于在整个采样过程中融入均匀噪声。如第 3 行和第 4 行所示，XDLM 和 UDLM 都在生成过程一开始就开始生成关键语义特征，具体来说就是狗的鼻子和眼睛。相比之下，MDLM 和 GIDD 采取了

更为保守的做法。它们几乎不会细化这些特定特征。因此，到最后一步[8/8], 时，MDLM 和 GIDD 仍无法确定鼻子和眼睛的正确结构，导致它们变形或缺失。然而，XDLM 很早就确立了这些特征，并持续加以细化，最终生成解剖结构正确且赏心悦目的图像。

这些模型在生成细腻纹理（例如狗毛）的能力上也存在显著差异。由于 MDLM 无法修正或细化先前生成的词元，因此生成的毛发纹理显得粗糙且缺乏连贯性。虽然 GIDD 生成的毛发比 MDLM 更细腻，但其表现受限于对时变噪声的依赖：有益的均匀噪声机制只有在最后一步才能有效发挥作用，从而限制了逐步改进的潜力。相比之下，XDLM 在所有步骤中都能持续细化毛发纹理，最终生成了其中最逼真、保真度最高的纹理。

最后但同样重要的是，这项比较表明，仅依赖均匀噪声（如 UDLM 所采用的方式）不足以实现最佳图像质量。尽管 UDLM 与 XDLM 一样能够较早定位特征，但最终生成的图像明显更加模糊。与成功兼顾结构连贯性和清晰、细致纹理的 XDLM 相比，这种清晰度不足导致细节丢失。

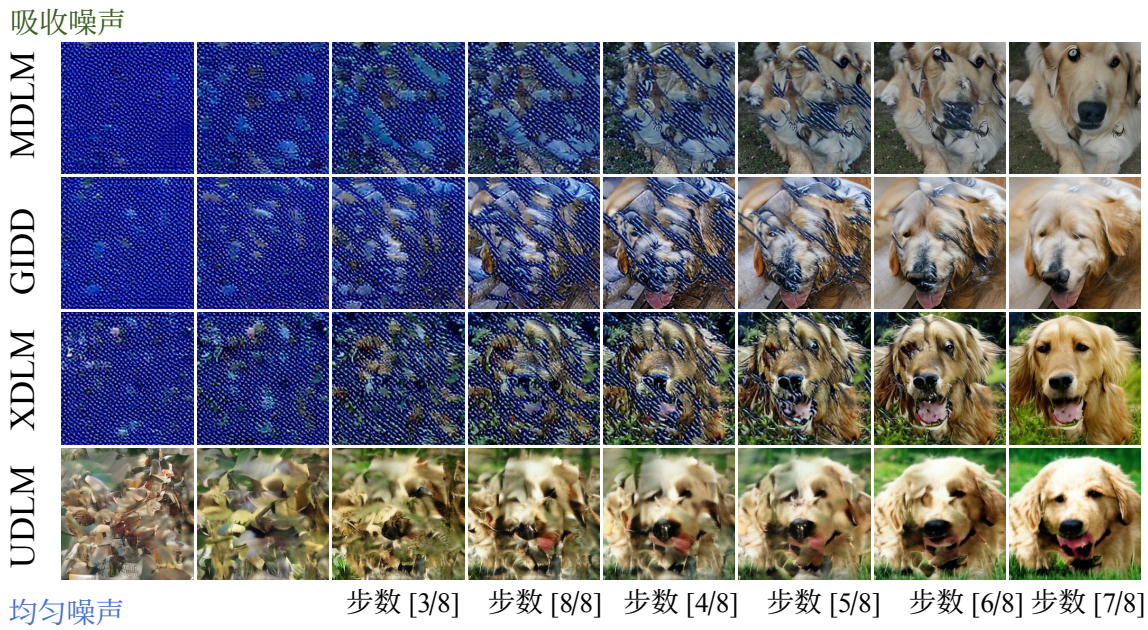

图7. **ImageNet-1K 上 8 步采样动态的视觉对比。** 该图对比了 MDLM、GIDD、XDLM（$k = 0.1$）以及 CFG尺度为 $= 2.0$ 的 UDLM 的生成轨迹。与基线模型相比，XDLM 展现出更优的结构连贯性与细节精细化能力，成功从噪声过渡到高保真度图像。

F. 实验详细配置

语言建模 遵循既有方法（Sahoo 等，2025；Schiff 等，2024；Sahoo 等，2024），我们使用 LM1B（Chelba 等，2013）和 OpenWebText（OWT）（Gokaslan 和 Cohen，2019）数据集，在标准基准测试上评估我们的模型。对于 LM1B，我们按照（Lou 等，2023；Sahoo 等，2024；2025）对数据集进行反词元化，并采用（Austin 等，2021；Sahoo 等，2025）的序列打包方式进行评估。序列使用 bert-base-uncased 分词器（Devlin 等，2019），以上下文长度 128 进行处理。对于 OWT，我们使用 GPT-2 分词器（Radford 等，2019），上下文长度为 1,024。序列以序列结束标记作为分隔符，打包至最大长度 1,024。由于 OWT 缺少标准验证集划分，我们将最后 10 万篇文档保留用于验证（Sahoo 等，2024）。

与先前工作一致（Sahoo 等人，2024年；Schiff 等人，2024年；Sahoo 等人，2025年），我们采用改进的扩散 Transformer（DiT）架构（Peebles 和 Xie，2023年），该架构改编自（Lou 等人，2023年）。所有模型均采用12层、12个注意力头、768的隐藏维度，以及维度为

128的时间步嵌入。丢弃率设为0.1，输入和输出词嵌入不共享。为确保公平比较，我们保留 AdaLN 分支，但将输入固定为零，而不是使用对数线性调度时间嵌入。

我们使用 AdamW（$\beta_1 = 0.9$、$\beta_2 = 0.999$、权重衰减 0）进行训练，全局批次大小为 512。学习率在前 2,500 步内逐渐升至 3×10^{-4}，之后保持不变。我们采用 0.9999 的指数移动平均（EMA）速率（Sahoo 等人，2024年；2025年）。

图像生成 我们在离散化的 CIFAR-10（Krizhevsky 等人，2009年）和 ImageNet-1K（Deng 等人，2009年）上进行评估。

在 ImageNet-1K 数据集上，我们使用 LlamaGen（Sun 等人，2024年）中的 VQ-VAE 分词器，将图像压缩为长度为 256 的序列。骨干架构与我们的语言建模 Transformer 相同，并加入类别条件嵌入以实现类别条件生成。训练超参数遵循语言建模配置，训练时长设为 0.5M 步。

对于 CIFAR-10，我们将原始 RGB 像素值视为离散词元（词汇表大小为 256），并将图像展平为长度为 $32 \times 32 \times 3$ 的序列，遵循（Schiff 等人，2024年）。我们采用 U-Net 骨干网络（Ronneberger 等人，2015年；Ho 等人，2020年），并使用离散化的截断逻辑斯蒂输出分布（Austin 等人，2021年；Sahoo 等人，2024年）。类别条件化通过将标签嵌入添加到时间步嵌入来实现（Dhariwal & Nichol，2021年）。该模型使用 AdamW（$\beta_1 = 0.9$，$\beta_2 = 0.999$，权重衰减 0）训练 300k 步，批次大小为 512。我们采用 5,000 步的学习率预热，峰值达到 2×10^{-4}，并保持指数移动平均（EMA）率为 0.9999。

大语言模型调优 我们通过在 LLaDA LLaDA（Nie 等人，2025年）上进行持续预训练，将 XDLM 扩展到大语言模型的语言建模任务中。LLaDA 最初通过 MDLM 训练。我们使用 FineWeb-Edu 数据集（Penedo 等人，2024年）中一个包含 100 亿词元的子集。训练配置采用 4,096 的序列长度和 512 的全局批次大小。学习率在 2×10^{-5} 经过 100 步预热后，在剩余的 500 步中保持不变。

除非另有说明，我们将混合比例设为 $k = 0.1$。所有实验均在配备 $8\times$ 台 Nvidia H800 GPU 的节点上进行。

G. 在 OWT 上训练的 XDLM 的零样本能力详情

为全面评估 XDLM 的泛化能力，我们在 OWT 上完成训练后，使用七个外部数据集评估零样本困惑度。本节将详细介绍实验设置，分析混合比例 k 的超参数敏感性，并对训练动态进行分解。

与先前工作保持一致（Sahoo 等人，2024年；2025年），我们的评估套件包含 AG News（Zhang 等人，2015）、LAMBADA（Paperno 等人，2016）、LM1B（Chelba 等人，2013年）、宾州树库（Marcus 等人，1993）、来自 arXiv 和 PubMed 的科学论文（Cohan 等人，2018）以及 WikiText（Merity 等人，2016）。验证困惑度通过采用蒙特卡洛采样计算负证据下界（ELBO）进行估计。我们在单个图形处理器上使用 128 的评估批次大小，并将所有文本预先打包为长度为 1024 的序列。为确保公平比较，我们采用与（Sahoo 等人，2024年；2025年）相同的采样配置。

我们研究混合比例 k 对模型性能的影响。表 7 和表8 分别展示了 OWT 上的验证困惑度以及外部基准测试上的零样本困惑度。

如表 7所示，XDLM 的性能与 k高度相关。当取值较小（例如 $k =$ 1e-3 和 $k = 0.1$）时，XDLM 的困惑度（PPL）分别为 23.495 和 24.097，与 MDLM（23.321）和 GIDD（23.136）相当。相反，随着 k 增至 0.9，性能下降至 25.731，接近 UDLM 基线（25.937）。

表 8中的零样本基准测试结果进一步印证了这一趋势。XDLM（$k = 0.1$）的平均困惑度为 54.110，与 MDLM（53.650）表现相近，并显著优于 UDLM（59.574）。这表明 XDLM 成功保留了掩码扩散范式的强大建模能力。

值得注意的是，XDLM 的优越性能并非仅体现在最终结果上，而是在整个训练过程中始终如一的特性。如图 8 所示，XDLM 的困惑度轨迹与 MDLM 高度一致，从 64.85 稳步下降至 54.11。相比之下，UDLM 曲线在所有训练步数中都明显更高。表 9 列出了以 100k 步为间隔统计的平均值零样本困惑度。数据证实，从早期阶段（100k 步）到收敛阶段（1M 步），XDLM（$k = 0.1$）都保持了与 MDLM 几乎相同的学习动态，凸显了其固有的

稳定性和效率 .

表7。 **训练步数为1M后在OWT上的验证困惑度。** 性能与混合比例 k 高度相关：较低的数值（$k \leq 0.1$）与强大的 MDLM 基线（23.321）保持一致，而较高的数值（$k \to 0.9$）则向 UDLM 基线（25.937）退化。

模型 MDLM		GIDD	XDLM(k=1e-3)	XDLM(k=0.1)	XDLM(k=0.5)	XDLM(k=0.9)	UDLM
PPL	23.321	23.136	23.495	24.097	24.818	25.731	25.937

表8. **在七个外部基准上的零样本困惑度评估。** XDLM（$k = 0.1$）展现出稳健的泛化能力，平均困惑度（54.110）与纯基于掩码的方法（MDLM/GIDD）相当，并显著优于 UDLM（59.574）。

数据集	AG News	LAMBADA	LM1B-GPT2	PTB	arXiv	PubMed	WikiText	平均值
MDLM	61.374	47.967	65.629	89.049	37.457	41.981	32.093	53.650
GIDD	60.607	47.811	65.898	86.911	39.019	42.634	30.809	53.384
XDLM(k=1e-3)	62.247	45.982	65.944	89.575	38.294	42.206	31.809	53.722
XDLM(k=0.1)	62.768	45.608	68.229	90.796	37.232	41.391	32.748	54.110
XDLM(k=0.5)	67.352	47.966	71.252	91.683	38.460	43.373	33.570	56.236
XDLM(k=0.9)	71.791	49.094	74.116	97.266	39.658	44.701	35.175	58.829
UDLM	69.402	51.272	75.572	95.986	42.671	47.181	34.933	59.574

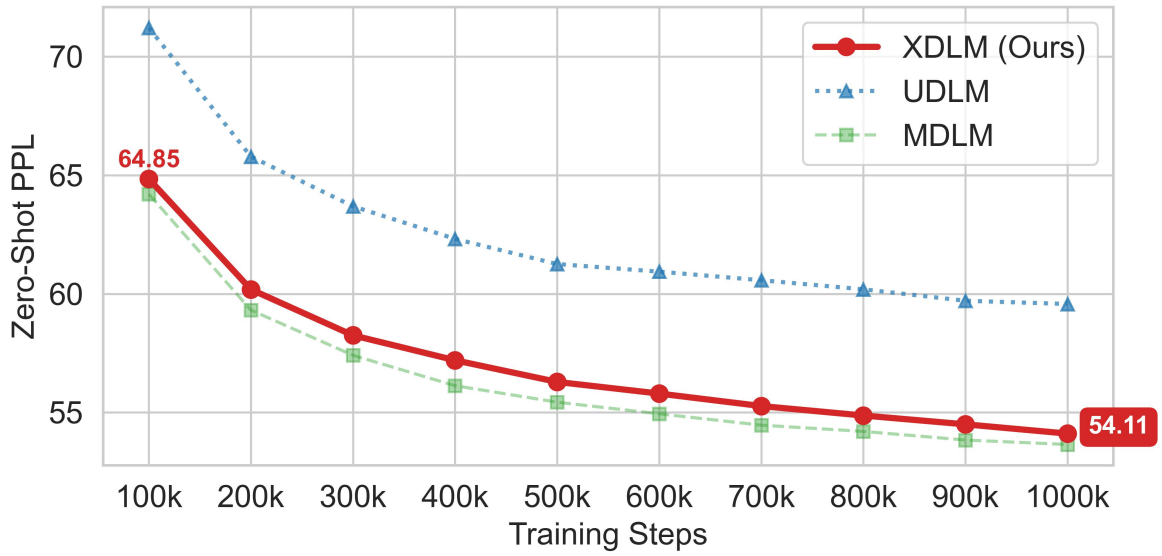

图 8. **训练全程的平均值零样本困惑度轨迹。** 与最终结果一致，XDLM（$k = 0.1$）从早期阶段（64.85）到收敛（54.11）都与 MDLM 基线高度吻合，而 UDLM 曲线在所有训练步数上始终明显更高。

表9. **训练全程零样本困惑度的详细演变**。 在每10万步间隔进行评估，数据证实XDLM（$k = 0.1$）保持着与MDLM几乎相同的学习动态，凸显了该方法从训练早期直至收敛的稳定性与效率。

训练步数	100k	200k	300k	400k	500k	600k	700k	800k	900k	1M
MDLM	64.209	59.315	57.411	56.129	55.434	54.943	54.466	54.201	53.839	53.650
GIDD	61.097	57.900	56.326	55.389	54.848	54.294	53.960	53.807	53.258	53.384
XDLM(k=1e-3)	63.748	59.097	57.319	56.374	55.576	54.937	54.577	54.336	53.977	53.722
XDLM(k=0.1)	64.847	60.187	58.253	57.201	56.288	55.797	55.271	54.872	54.500	54.110
XDLM(k=0.5)	68.074	62.524	60.489	59.353	58.396	57.881	57.437	56.976	56.530	56.236
XDLM(k=0.9)	70.264	65.384	63.147	61.758	60.977	60.263	59.723	59.438	58.915	58.829
UDLM	71.220	65.789	63.692	62.320	61.258	60.937	60.577	60.192	59.711	59.574

H. 详细的语言生成结果

在本节中，我们给出语言生成实验的全面数值结果，这些实验最初在第4.2节中讨论。为系统分析吸收式（类 MDLM）与均匀式（类 UDLM）扩散项过程之间的权衡，我们在不同混合比例 $k \in \{10^{-3}, 0.1, 0.5, 0.9\}$下评估 XDLM。对于生成，我们采用祖先采样设置（Sahoo 等人，2024年；Schiff 等人，2024年；Sahoo 等人，2025年），通过 Gumbel 噪声从估计后验进行并行词元采样。所得性能基于样本质量进行评估，由 GPT-2 Large 困惑度（困惑度 ↓）量化，多样性则由词元熵（↑）衡量。

表10 给出了 OWT 数据集上的全面评估。数据定量证实， k 决定了模型在不同采样步数下的性能，并有效地在 UDLM 与 MDLM 的行为之间插值。在少步数区间（例如 8 至 32 个采样步数）中，增大 k 可显著提升性能，使 XDLM 在效率上与 UDLM 对齐。例如，仅用 8 个采样步数时，XDLM （ $k = 0.9$ ）得到 189.750 的低困惑度，与 UDLM（183.991）极具竞争力，且远优于基于掩码的基线方法 MDLM（711.382）和 GIDD（654.841）。相反，在多步数区间（例如 512 至 1024 个采样步数）中，保留类掩码特性（更低的 k）对实现高保真生成仍然有利。XDLM（$k = 0.1$）展现出最优平衡。当采样预算有限时，它相较 MDLM 有大幅提升；而当可用采样步数更多时，它仍能达到极低困惑度（1024 步时为 52.609），且不牺牲熵所体现的多样性。

LM1B 数据集上的结果（详见表 11）也印证了这些发现。我们观察到一个一致的趋势：当采样步数较少时，较高的 k 占据主导。在 4 个采样步数下，XDLM（$k = 0.5$）的困惑度为 232.702，与 UDLM（232.403）基本持平，并显著优于 MDLM（377.177）。随着采样步数增加，性能差距逐渐缩小，但 XDLM 在整个范围内始终表现稳健。总的来说，这些结果表明，通过调整 k，XDLM 弥合了两种范式之间的差距：采样步数较少时，它兼具均匀噪声的效率；步数较多时，则保留了掩码方法实现高质量结果的潜力。

表 *10.* **OWT 上的质量（困惑度）与多样性（熵）**。结果表明，最佳混合比例 k 取决于采样预算。在少步数区间（例如 8 步），较高的 k 对于提高效率至关重要，使 XDLM 的性能能够媲美 UDLM。相反，在多步数区间，较低的 k 更有利于实现高保真度。值得注意的是，XDLM（$k = 0.1$）达到了最佳平衡：在步数较少时显著优于 MDLM，在步数较多时则获得更低的困惑度。

模型	8步	16步	32步	步数 64	步数 128	步数 256	步数 512	步数 1024
				困惑度 ↓				
MDLM	711.382	288.157	156.576	106.098	80.198	64.376	51.756	41.545
GIDD	654.841	273.829	151.331	102.558	77.548	62.077	49.822	39.981
XDLM(k=1e-3)	613.647	253.897	146.567	104.090	80.301	65.245	52.650	42.797
XDLM(k=0.1)	301.640	161.724	114.795	94.837	83.205	73.824	63.888	52.609
XDLM(k=0.5)	218.572	127.335	98.817	86.956	81.051	77.155	73.503	66.136
XDLM(k=0.9)	189.750	117.234	93.239	84.068	80.614	79.456	78.707	78.238
UDLM	183.991	111.237	88.632	79.509	76.611	74.019	73.345	106.666
				熵 ↑				
MDLM	5.872	5.767	5.687	5.614	5.543	5.472	5.387	5.287
GIDD	5.832	5.748	5.675	5.605	5.534	5.462	5.370	5.269
XDLM(k=1e-3)	5.837	5.746	5.673	5.613	5.546	5.475	5.393	5.298
XDLM(k=0.1)	5.662	5.630	5.600	5.573	5.548	5.516	5.467	5.390
XDLM(k=0.5)	5.574	5.565	5.555	5.547	5.538	5.525	5.514	5.476
XDLM(k=0.9)	5.546	5.547	5.548	5.548	5.548	5.547	5.544	5.542
UDLM	5.536	5.539	5.543	5.535	5.536	5.531	5.524	5.769

表*11.* **LM1B 上的质量（困惑度）与多样性（熵）**。这与 OWT 的发现相印证：当采样预算严格受限时，较高的 k 占据主导地位。在 4 步时，XDLM（$k = 0.5$）的表现与 UDLM 基本相当。在整个采样步数范围内，XDLM 弥合了两种范式之间的差距：预算较小时，它兼具均匀噪声的效率；预算较大时，则保留了掩码的高质量潜力。

模型	第 4 步	8步	16步	32步	64步	128步
			困惑度 ↓			
MDLM	377.177	215.579	152.939	124.731	112.049	102.142
GIDD	392.002	215.873	151.824	124.014	110.247	101.211
XDLM(k=1e-3)	343.246	199.457	146.612	124.444	111.960	103.544
XDLM(k=0.1)	246.134	155.487	123.129	113.006	105.509	101.983
XDLM(k=0.5)	232.702	141.121	116.063	106.182	103.419	100.754
UDLM	232.403	136.914	110.575	102.072	98.571	96.385
			熵 ↑			
MDLM	4.405	4.379	4.366	4.357	4.353	4.347
GIDD	4.401	4.373	4.361	4.359	4.354	4.348
XDLM(k=1e-3)	4.382	4.363	4.357	4.358	4.354	4.351
XDLM(k=0.1)	4.326	4.329	4.334	4.342	4.345	4.346
XDLM(k=0.5)	4.309	4.315	4.325	4.335	4.340	4.342
UDLM	4.303	4.305	4.320	4.327	4.334	4.336

I. 详细图像生成结果

我们对 XDLM 在 ImageNet-1K 和 CIFAR-10 上的图像生成能力进行了扩展分析，重点考察混合比例参数 k的影响。根据我们的方法论定义，k 决定了基于掩码的过程与均匀噪声过程之间的平衡。较高的 k 会提高均匀噪声所占的比例（实际效果趋近于 UDLM 设置），而较低的 k 则会保留更多类掩码特性。表 12、13、14、15 中的结果从实证角度验证了这种插值的作用。我们观察到两种关键表现：首先，XDLM 持续收敛至

UDLM 的行为随 k 增加而变化；其次，最佳性能往往出现在插值区域（$0 < k < 1$），而非两个极端（纯 MDLM 或 UDLM）。

表 12 和 13 详细展示了 ImageNet-1K 上的性能，为最优 k 取决于生成设置这一点提供了有力证据。在标准条件设置下（表 12），均衡混合效果更佳：XDLM 在 $k = 0.5$下以 16 步达到 23.417 的 FID，显著优于低混合变体（$k = 10^{-3}$）和专用 UDLM 基线（26.242）。然而，应用无分类器引导（CFG）后（表 13），最优工作点会发生变化。此时，$k = 0.1$ 成为“最佳平衡点”，FID 达到 8.625，优于纯 UDLM（8.980）和 $k = 0.5$ 变体（8.790）。这些结果凸显了 XDLM 的优势。通过探索掩码与均匀噪声之间的插值空间，我们能够找到优于两种单独基线的配置。

表 14 和 15进一步说明了由以下因素控制的行为转变：k 在 CIFAR-10 上。在 UDLM 展现出明显优势的这一基准测试中，我们观察到，随着 k 增大，生成质量呈现清晰的单调提升。例如，在标准条件下，将 k 从 10^{-3} 提高到 0.5 可将 32 步 FID 从 164.040 降至 56.299，大幅缩小与 UDLM（41.027）之间的差距。使用 CFG 时也有类似趋势，此时 $k = 0.5$接近 UDLM 基准。总而言之，这些结果证实 XDLM 成功弥合了不同领域之间的差距。至关重要的是，这表明不存在适用于所有场景的单一固定 k；相反，要在不同数据集和引导机制下实现性能最大化，必须能够灵活地在区间 k 内调整$(0,1)$。

表 *12.* **ImageNet-1K 上 FID（↓）和 IS（↑）的比较。** 对于图像生成，XDLM 展现出强劲的竞争力，性能与专用 UDLM 模型高度相当。值得注意的是，在 16 个生成步数下，XDLM 的表现优于包括 UDLM 在内的所有基线，证实了其在文本领域之外同样具备稳健的性能。

模型	*FID*↓				*IS*↑			
	4步	**8步**	**16步**	**32步**	**4步**	**8步**	**16步**	**32步**
MDLM	80.752	47.732	28.785	18.928	16.287	29.178	44.656	57.803
GIDD	86.842	54.933	35.403	24.588	14.559	24.297	35.698	46.376
XDLM(k=1e-3)	81.992	52.590	35.046	25.637	15.597	24.813	35.097	44.032
XDLM(k=0.1)	54.085	34.109	25.774	22.187	24.829	36.964	43.903	48.118
XDLM(k=0.5)	45.808	29.362	23.417	21.486	29.310	40.482	46.620	48.911
UDLM	49.861	30.144	26.242	25.661	27.049	38.832	41.801	41.850

表*13.* **在ImageNet-1K上使用CFG（尺度 $= 2.0$）的性能。**引导的应用会改变最优工作点。在此，$k = 0.1$ 成为“最佳平衡点”，在16步时FID达到8.625。该配置同时超越纯UDLM基线（8.980）与更高混合度变体（$k = 0.5$），表明针对特定引导机制调节 k 的重要性。

模型	*FID*↓				*IS*↑			
	4步	**8步**	**16步**	**32步**	**4步**	**8步**	**16步**	**32步**
MDLM	33.468	11.144	6.725	7.485	54.740	119.150	172.664	203.722
GIDD	41.000	15.151	7.076	6.314	44.084	95.789	148.148	181.017
XDLM(k=1e-3)	32.876	12.872	7.612	7.588	53.351	102.601	144.019	167.660
XDLM(k=0.1)	13.550	8.956	8.625	8.914	107.403	148.723	165.916	171.046
XDLM(k=0.5)	11.945	9.038	8.790	8.939	114.135	145.718	156.339	159.829
UDLM	14.055	9.718	8.980	8.650	97.859	123.582	132.099	132.108

表*14*。**CIFAR-10 上的类别条件生成**。在 UDLM 具有明显优势的这一基准测试中，我们观察到，随着 k 增加，XDLM 的质量持续提升。将 k 从 10^{-3} 提高到 0.5，大幅缩小了与 UDLM 的性能差距（例如，32步 FID 从 164.040 提升至 56.299），证实了较高的 k 能有效使 XDLM 与均匀噪声动态保持一致。

模型	*FID*↓ 32步	128步	512步	*IS*↑ 32步	128步	512步
MDLM	211.983	69.151	33.961	2.685	5.615	6.801
GIDD	193.487	72.579	41.809	2.737	5.176	6.209
XDLM(k=1e-3)	164.040	50.638	29.079	3.427	6.223	7.079
XDLM(k=0.1)	77.985	48.702	43.210	5.316	6.284	6.513
XDLM(k=0.5)	56.299	37.307	33.030	6.077	6.801	6.956
UDLM	41.027	27.822	25.144	6.892	7.345	7.372

表*15*。 **使用 CFG 时 XDLM 与基线在图像生成方面的性能** $scale = 2.0$**在 CIFAR-10 上**。 应用 CFG 后，所有基于掩码的方法（MDLM、GIDD、XDLM）都取得了显著改进。总体趋势与未使用 CFG 时的结果一致。

模型	*FID*↓ 32步	128步	512步	*IS*↑ 32步	步骤 128	步骤 512
MDLM	181.453	51.113	22.762	3.329	6.969	8.078
GIDD	179.916	60.443	30.812	3.054	6.124	7.216
XDLM(k=1e-3)	130.106	35.980	19.877	4.559	7.575	8.186
XDLM(k=0.1)	59.600	36.889	32.444	6.699	7.596	7.670
XDLM(k=0.5)	40.850	27.445	24.445	7.560	7.881	7.893
UDLM	30.328	20.720	18.592	8.098	8.392	8.433

J. LLaDA 持续预训练结果详述

为在 LLaDA 框架内通过实证验证 XDLM 目标的有效性，我们使用 LLaDA 8B Base 检查点开展了持续预训练实验。我们将 XDLM 建模定义集成到 transformers 代码库中，并在 FineWeb-Edu 数据集的 100 亿词元采样上进行训练（Penedo 等人，2024年）。训练过程共进行 600 步，全局批次大小为 512，序列长度为 4,096。学习率在前 100 步内预热至 2×10^{-5}，随后 500 步保持不变；XDLM 混合比例设为0.1。所有实验均在一台配备 $8\times$ Nvidia H800 GPU 的计算节点上进行。为严格地将性能提升归因于架构范式，而非混杂变量，我们建立了两个可靠的基线：LLaDA-MDLM，即对照组，模型在相同时间内使用标准掩码扩散语言模型目标进行持续预训练；以及 LLaDA-XDLM-infer，即消融实验，在不进一步训练的情况下，将 XDLM 采样策略应用于基础模型。

为了在这一架构中部署 XDLM 采样，我们扩展了 LLaDA 中使用的标准低置信度重新掩码策略。算法1 详细说明了生成逻辑，重点介绍由混合比例 k控制的条件块中的创新贡献。标准 LLaDA 采样通过单调降低不确定性来运行，只优先填充已有的 `[MASK]` 词元，而我们的实现引入了细化分支。当 $k > 0.0$时，算法会计算“细化_优先级”：找出当前未被掩码（“is_masked”）但模型预测了与之冲突的词元 ID（“pred_tokens != token_ids”）且其置信度高于当前词元置信度的词元。这些词元会与标准空掩码一同被标记到“update_mask”中。这一机制从根本上将生成过程从纯粹的增量过程转变为动态纠错循环，使模型能够以随机方式“改变主意”，并在去噪轨迹中修正先前的输出。

表 16中汇总的实证结果表明，该公式在推理（GSM8K、MATH、BBH）和代码生成（HumanEval、MBPP）基准测试中均具有优越性，尤其是在低计算量场景下。在 32 个采样步数下，LLaDA-XDLM 的表现显著优于所有基线。最明显的优势体现在 MBPP 代码生成任务中：LLaDA-XDLM 的得分达到 15.00，几乎是基础 LLaDA（6.80）和持续训练的 MDLM 对照模型（4.40）表现的两倍。对 MBPP 失败模式的细粒度分析显示，这一改进

源于生成保真度的根本提升；LLaDA-XDLM 大幅降低了“失败”案例的发生率，这些案例被定义为无法编译或存在语法错误的代码，从 LLaDA 的 429 例降至 304 例。通过成功将这些此前无效的尝试转化为“通过”的输出（75 例，而非 34 例），该模型证实，XDLM 目标与细化采样策略相结合，能够提供必要的结构连贯性，从而自行纠正原本会导致编译失败的语法错误。当采样步数增加到 64、128 和 256（总生成长度）时，LLaDA-XDLM 仍保持竞争优势，这验证了性能提升直接源于 XDLM 训练过程中学习到的内部表示，而非仅仅是采样算法造成的结果。

值得注意的是，我们观察到持续预训练基线（LLaDA-MDLM）相较于原始LLaDA出现了性能下降，这很可能是由于微调数据与原始预训练语料库之间的分布偏移所致。因此，LLaDA-XDLM优于LLaDA-MDLM的表现证实，我们的收益源于XDLM形式的有效性，而非仅仅来自额外的600个训练步数。

此外，LLaDA-XDLM显著优于消融基线LLaDA-XDLM-infer，这表明改进内在于所学模型，而非采样技巧造成的假象。

算法 1 LLaDA-XDLM生成策略

```
def 生成(步数, 序列_长度, 掩码_id, topk_吸收, topk_均匀, k=0.1):"""
  参数:
    步数（整数）：采样的总步数。
    序列_长度（整数）：要生成的目标序列长度。
    掩码_id（整数）：[MASK] 词元的词汇表索引。
    topk_吸收（整数列表）：每个步数要解码的词元数量计划。
    topk_均匀（整数列表）：每个步数要细化的词元数量计划。
    k（浮点数）：混合比例。
  返回:
      词元_ID（torch.Tensor）：生成的 token ID 序列，形状为 (seq_length,)。"""
    词元_ID = torch.full((seq_length,), fill_value=mask_id, dtype=torch.long)
    对于 i 在 range(steps):
        is_被掩码 = (词元_ID == 掩码_id)
        logits = 模型(token_ID).logits
        logits_加上_噪声 = 添加_Gumbel_噪声(logits)
        预测_词元 = torch.argmax(logits_加上_噪声, dim=-1)
        概率 = torch.softmax(logits, dim=-1)
        预测_置信度 = torch.squeeze(torch.gather(probs, dim=-1, index=torch.unsqueeze(pred_词元, -1)), -1)
        mask_优先级 = torch.where(is_masked, pred_confidence, -np.inf)

        if k > 0.0:
            当前_置信度 = torch.squeeze(torch.gather(probs, dim=-1, index=torch.unsqueeze(token_ids, -1)), -1)
            可以_被_细化 = （~已_掩码）&（预测_词元! = 词元_标识）&（预测_置信度 >= 当前_置信度）
            细化_优先级 = torch.where(可以_被_细化, 预测_置信度, -np.inf)

        更新_掩码 = torch.zeros_like(预测_词元, dtype=torch.bool) _, 最高_掩码_索引 =
        torch.topk(掩码_优先级, k=topk_吸收 [i]) 更新_掩码 [最高_掩码_索引] =True


        如果 k > 0.0:
            _, 最高_细化_索引 = torch.topk(细化_优先级, k=topk_均匀 [i])
            更新_掩码 [最高_细化_索引] = True

        词元_编号 [更新_掩码] = 预测_词元 [更新_掩码]

    返回 词元 ID_
```

表 *16.* **LLaDA-XDLM 在推理和代码生成基准测试上的评估**。我们将所提出的方法与基础模型 LLaDA、仅推理消融（LLaDA-XDLM-infer）以及 MDLM 对照模型（LLaDA-MDLM）进行比较，并考察不同采样步数下的表现。MBPP 得分按失败模式进行分解，以展示结构正确性。

数据集	BBH	GSM8K	MATH	HumanEval	MBPP			
					得分	失败	错误	Pass
32 步数 / 256 词元								
LLaDA	42.68	24.49	4.80	5.49	6.80	429	37	34
LLaDA-XDLM-infer	41.74	24.26	4.86	1.22	5.40	438	35	27
LLaDA-MDLM	43.21	24.41	2.70	6.71	4.40	447	31	22
LLaDA-XDLM	42.76	29.26	4.72	10.98	15.00	304	121	75
64 步数 / 256 词元								
LLaDA	47.16	57.16	16.10	12.80	17.80	328	83	89
LLaDA-XDLM-infer	46.76	53.75	16.12	4.88	17.60	328	84	88
LLaDA-MDLM	48.32	54.97	13.42	14.63	13.40	348	84	67
LLaDA-XDLM	45.61	57.85	16.48	15.85	23.60	197	182	118
128 步数 / 256 词元								
LLaDA	47.49	67.93	26.12	24.39	27.60	161	199	138
LLaDA-XDLM-infer	47.89	66.41	26.70	8.54	27.40	187	174	137
LLaDA-MDLM	48.95	68.61	24.34	25.00	30.00	169	178	150
LLaDA-XDLM	46.41	68.01	25.06	25.00	30.80	130	215	154
256 步数 / 256 词元								
LLaDA	47.77	72.25	29.98	33.54	40.40	48	248	202
LLaDA-XDLM-infer	47.93	71.34	29.34	6.71	37.00	74	238	185
LLaDA-MDLM	49.84	72.93	29.70	32.32	37.60	51	260	188
LLaDA-XDLM	46.07	73.16	29.22	31.71	34.60	90	236	173

K. 详细计算效率分析

我们通过在三种不同场景下测量速度和内存消耗，评估所提出的 XDLM 与基线模型在计算效率方面的表现：仅前向 场景（代表用于困惑度估计的 NELBO 计算）、前向-反向 场景（用于训练）以及采样 场景（用于生成）。对于前向和前向-反向评估，我们使用批次大小为 32、序列长度为 1024 的随机生成序列，以模拟真实数据处理。为确保测量结果稳健，我们先对模型进行 10 次预热迭代，并报告随后 100 次运行所计算的平均值。对于采样场景，我们采用标准祖先采样，在 32 个步数内生成长度为 1024 的序列。评估批次大小设为 32。模型先进行 1 次预热迭代，并报告随后 10 次运行所计算的平均值。

如表 17 所示，MDLM 在所有设置下均展现出最高的吞吐量和最低的内存消耗，这一结果并不意外，因为其吸收噪声核十分简单。然而，在采用均匀噪声分布的模型中，XDLM 在前向、前向-反向和采样模式下的吞吐量均最高。这一表现直接归功于我们采用的标量化重构高效采样与训练策略，该策略避免了开销巨大的矩阵运算。相比之下，UDLM 在前向和训练模式下的吞吐量排名第三，但在采样模式下的吞吐量却是对比组中最低的。反之，GIDD 的采样吞吐量排名第三，但 NELBO 计算表现不佳，其前向吞吐量仅为 XDLM 的一半。内存消耗方面也呈现出类似趋势。MDLM 整体上最为轻量，而在噪声核包含均匀噪声分量的方法中，XDLM 排名第一，有效避免了 GIDD 和 UDLM 所呈现的高内存开销。

表 *17.* **计算效率比较**。 我们报告了 MDLM、GIDD、XDLM 和 UDLM 的吞吐量（词元/秒， ↑）和峰值内存使用量（GB， ↓）。这些指标涵盖三种场景：前向 （推理/困惑度估计）、前向-反向 （训练）和采样 （生成）。借助标量重构策略，XDLM 实现了极具竞争力的吞吐量和内存效率，在大多数指标上优于采用均匀噪声核的其他基线（GIDD、UDLM）。

模型	吞吐量（词元/秒） ↑			内存（GB） ↓		
	前向	前向-反向	采样	前向	前向-反向	采样
MDLM	424294	141990	8789	6.285	35.354	18.848
GIDD	199516	95395	6336	25.131	51.060	40.856
XDLM (k=0.1)	396398	137372	7108	18.850	41.634	31.414
UDLM	370952	142276	2882	18.850	44.777	59.683

L. 详细训练动态

本节提供第 4.3 节所讨论的训练动态分析的完整数值数据，并补充图 4中所示的可视化结果。这些表格通过给出训练过程中困惑度、熵、FID 和 IS（初始分数）的精确指标值，补充了可视化分析。

表18 详细列出了LM1B基准在100万训练步数过程中的训练动态。各模型使用固定预算的128个采样步数进行评估。如正文所述，尽管MDLM在初始阶段表现更优，UDLM在后期训练阶段展现出强劲的扩展能力，并在100万步时取得各基线中最低的困惑度（96.385）。此外，该表还详细说明了改变混合比例 k 在我们提出的XDLM方法中所带来的影响。

作为对文本生成结果的补充，表19 给出了ImageNet-1K上图像生成的定量训练动态。评估在50万训练步数内进行，采用固定的16个采样步数预算。数值数据印证了视觉趋势：XDLM确立并保持了决定性的性能优势。值得注意的是，使用 $k = 0.5$的XDLM在最终检查点取得了所报告的最低FID 23.417和最高IS（初始分数）46.620，并持续优于MDLM和GIDD基线。尽管UDLM的起始FID较高，但它迅速收敛至具有竞争力的分数（26.242）；然而，XDLM在大部分训练阶段中始终是领先模型。

表*18.* **LM1B基准上的定量训练动态，补充图 4 中的可视化分析**。 我们报告在 1M 训练步数内的困惑度（↓）和熵（↑），采样步数固定为 128。该表详细展示了不同混合比例 k 下 XDLM 相较于 MDLM、GIDD 和 UDLM 基线的表现。

训练步数	100k	200k	300k	400k	500k	600k	700k	800k	900k	1M
困惑度 ↓										
MDLM	113.218	107.864	106.735	105.839	104.594	104.106	103.559	103.174	102.769	102.142
GIDD	107.366	103.473	103.386	101.854	102.367	101.745	101.904	101.957	101.490	101.211
XDLM(k=1e-3)	112.628	108.381	106.406	104.913	105.156	103.817	103.624	102.731	103.638	103.544
XDLM(k=0.1)	119.632	110.803	109.017	106.265	105.428	104.572	103.645	103.348	102.223	101.983
XDLM(k=0.5)	117.846	111.024	108.130	105.955	104.953	103.048	102.464	101.040	102.015	100.754
UDLM	116.021	107.800	104.508	102.466	101.480	99.958	99.660	98.147	97.503	96.385
熵 ↑										
MDLM	4.350	4.348	4.350	4.347	4.349	4.349	4.348	4.349	4.349	4.347
GIDD	4.352	4.350	4.351	4.349	4.351	4.349	4.350	4.350	4.351	4.348
XDLM(k=1e-3)	4.351	4.347	4.349	4.347	4.350	4.347	4.347	4.349	4.351	4.351
XDLM(k=0.1)	4.353	4.347	4.347	4.344	4.344	4.344	4.346	4.344	4.344	4.346
XDLM(k=0.5)	4.350	4.346	4.345	4.344	4.342	4.343	4.342	4.343	4.344	4.342
UDLM	4.349	4.343	4.341	4.340	4.338	4.336	4.338	4.336	4.335	4.336

表 *19.* **ImageNet-1K 生成的定量训练动态**，对应图中的曲线4。我们报告了 500k 训练步数内的 FID（↓）和 IS（↑），并采用固定的 16 个采样步数预算。数值数据印证了视觉趋势：XDLM 确立并保持了显著的性能优势。

训练步数	*FID*↓					*IS*↑				
	100k	**200k**	**300k**	**400k**	**500k**	**100k**	**200k**	**300k**	**400k**	**500k**
MDLM	40.060	33.723	31.090	29.436	28.785	28.978	36.136	39.804	42.893	44.656
GIDD	42.152	36.878	35.200	35.677	35.403	27.337	33.042	35.544	35.173	35.698
XDLM(k=1e-3)	45.381	38.967	36.061	36.287	35.046	24.967	30.512	33.948	34.031	35.097
XDLM(k=0.1)	38.358	31.464	27.524	26.853	25.774	27.949	35.839	41.150	42.222	43.903
XDLM(k=0.5)	37.567	30.098	26.687	23.790	23.417	27.854	35.682	40.724	45.976	46.620
UDLM	43.316	34.525	30.381	28.086	26.242	23.619	31.188	35.963	38.588	41.801