连续扩散语言模型（Continuous Diffusion Language Models）

黄京

Aug 30, 2026

1 从图像到文本：扩散模型的跨模态尝试

扩散模型在图像生成领域取得的成功已无需赘述，其核心思想是通过逐步添加噪声将数据分布映射到标准高斯，再通过学习逆过程实现高质量样本生成。然而，将这一范式迁移到文本领域面临根本性挑战：文本由离散符号构成，而扩散过程天然依赖连续空间。这促使研究者探索将离散 token 映射到连续嵌入空间的方案，形成了连续扩散语言模型这一新兴方向。该类模型试图在保留扩散模型全局建模优势的同时，解决文本生成的离散性约束问题。

2 扩散模型的数学框架

在连续扩散语言模型中，前向加噪过程首先将 one-hot 形式的 token 转换为稠密嵌入向量。给定原始嵌入 x_0，扩散过程定义为 $q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$，其中 $\bar{\alpha}_t$ 控制噪声强度随时间步递增。这一设计使得在任意时刻 t，我们都能从原始数据直接采样得到加噪版本，无需迭代累积。

反向去噪过程则试图学习参数化的条件分布 $p_\theta(x_{t-1}|x_t) \approx \mathcal{N}(\mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$。实践中通常采用简化目标：直接预测添加的噪声 ε 或直接预测原始数据 x_0。前者对应于最小化 $\mathbb{E}_{t,x_0,\varepsilon}\|\varepsilon - \varepsilon_\theta(x_t,t)\|^2$，后者则优化重建损失。两种目标在不同噪声调度下表现出互补特性，噪声预测在高噪声区域更稳定，而 x_0 预测在低噪声区域重建更精确。

3 模型架构与实现细节

连续扩散语言模型的嵌入层将离散词表映射为可学习的连续向量空间。不同于 VQ-VAE 的硬量化策略，CDLM 保持嵌入的连续性，允许梯度在嵌入空间自由流动。嵌入维度通常设置为 768 – 1024，与 Transformer 主干的隐藏维度保持一致。部分工作探索了冻结预训练词嵌入的可能性，但实验表明可学习嵌入在下游任务上通常获得 2 – 3 个点的 PPL 提升。

主干网络采用双向 Transformer encoder，允许每个位置在所有时间步都能看到完整上下文。时间步信息通过 Sinusoidal 编码或可学习嵌入注入，常用 AdaLN（Adaptive Layer Normalization）机制实现：对每个 Transformer 层计算 $\gamma(t)$ 和 $\beta(t)$，对隐藏状态进行 scale-shift 调制。代码实现中，时间步嵌入通常先经过两层 MLP 映射到与隐藏维度相同的空间，再与层归一化参数融合。

```python
class TimeEmbedding(nn.Module):
    def __init__(self, dim):
        super().__init__()
        self.dim = dim
        # Sinusoidal 位置编码
        self.proj = nn.Sequential(
            nn.Linear(dim, dim * 4),
            nn.SiLU(),
            nn.Linear(dim * 4, dim)
        )

    def forward(self, t):
        # t: [batch_size]
        half = self.dim // 2
        freqs = torch.exp(-torch.arange(half, device=t.device) *
                    math.log(10000) / (half - 1))
        args = t[:, None].float() * freqs[None]
        emb = torch.cat([torch.cos(args), torch.sin(args)], dim=-1)
        return self.proj(emb)
```

上述代码实现了时间步的 Sinusoidal 编码与 MLP 投影。输入时间步 t 首先扩展为正弦余弦特征，再通过两层非线性变换得到与模型隐藏维度匹配的向量。该向量随后在 AdaLN 中用于调节 Transformer 层的归一化参数，实现条件生成能力。

4 训练策略与工程优化

时间步采样策略对模型收敛至关重要。均匀采样在 $[1, T]$ 区间内随机选择 t，但由于不同 t 对应的去噪难度差异巨大，带权采样通常更有效。常见做法是按照噪声调度函数 β_t 的密度进行重要性采样，使模型在高噪声和低噪声区域都能获得充分训练。实践中，权重函数常设计为 $w(t) \propto \beta_t / \alpha_t$，平衡重建误差与梯度方差。

序列长度外推是长文本生成的关键瓶颈。标准绝对位置编码难以泛化到训练长度之外，相对位置编码（如 ALiBi）或外推式编码（如位置插值）成为必要选择。ALiBi 通过在注意力分数上添加线性偏置，允许模型在推理时处理更长序列而无需重新训练。实验表明，采用 ALiBi 的 CDLM 模型可以将有效上下文长度从 512 扩展至 2048，同时保持困惑度基本不变。

5 推理加速与采样策略

标准 DDPM 采样需要数百至上千步迭代，远慢于自回归模型的一次前向。DDIM 通过将扩散过程重新参数化为非马尔可夫链，允许用数十步甚至数步完成高质量采样。其核心公式为 $x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\varepsilon_\theta(x_t, t)$，其中 σ_t 控制随机性大小。实践中，设置 $\sigma_t = 0$ 可获得确定性采样，进一步加速推理。

动态 early-stopping 根据当前去噪质量自适应调整步数。当预测的 x_0 与前一步的重建差异低于阈值时，可提前终止扩散过程。该策略在低噪声区域特别有效，因为此时模型已接近最终输出，继续迭代收益递减。结合键值缓存机制，推理吞吐量可提升 3 – 5 倍。

6 条件控制机制

Classifier-Free Guidance（CFG）是 CDLM 中最常用的条件控制方法。训练时随机以一定概率丢弃条件信息，强制模型同时学习条件与无条件分布。推理时通过插值公式 $\tilde{\varepsilon}_\theta = \varepsilon_\theta(x_t, t, \emptyset) + s \cdot (\varepsilon_\theta(x_t, t, c) - \varepsilon_\theta(x_t, t, \emptyset))$ 实现可调强度控制，其中 s 为 guidance scale。较大的 s 增强条件遵循度但降低多样性，较小的 s 则相反。
跨模态条件通过预训练编码器将外部信号映射到与文本嵌入相同的空间。CLIP 图像编码器输出可直接作为扩散模型的条件向量，实现图生文任务。结构化条件（如表格、JSON）则需要专门的编码器，将键值对序列化为连续表示，再与时间步嵌入融合。实验表明，条件注入位置（早期 vs. 晚期层）对生成质量影响显著，早期注入通常获得更强的条件遵循能力。

7 长文本建模挑战

标准 Transformer 的平方复杂度限制了 CDLM 处理长文本的能力。层次化扩散通过先在句子级别进行扩散，再在词元级别细化，有效降低了计算复杂度。第一阶段在粗粒度序列上运行，第二阶段以第一阶段输出为条件进行细粒度去噪。状态空间模型（如 Mamba、RetNet）与扩散的结合是另一前沿方向，这些模型提供线性复杂度且保持长程依赖建模能力，适合作为扩散主干的替代架构。

8 理论分析与开放问题

尽管扩散模型在图像 FID 上表现优异，但在文本 PPL 上仍落后于自回归模型。这一差距源于信息瓶颈：连续嵌入空间需要通过有限维度表示离散符号的概率分布，导致表示坍缩。理论分析表明，最优噪声调度应与数据流形曲率相关，但实践中常用线性或余弦调度仅为近似。收敛速度分析显示，扩散模型的 ELBO 与对数似然之间存在间隙，该间隙随序列长度增加而扩大。
与能量模型、流模型的统一视角揭示了扩散模型的本质：它们都是通过学习从噪声到数据的映射来隐式建模数据分布。能量模型直接定义概率密度，流模型学习可逆变换，而扩散模型则通过马尔可夫链逐步转换。三者在数学上可相互转化，提供了从不同角度理解生成过程的可能。

9 实际应用与生态现状

开源实现中，Diffusion-LM 和 CD-LM 提供了 PyTorch 参考代码，HuggingFace Diffusers 计划集成扩散语言模型支持。学术预训练权重覆盖 110M 至 1B 参数规模，社区微调脚本支持在特定领域数据上继续训练。快速上手示例显示，通过十余行代码即可加载预训练模型并生成句子，CFG scale 参数可直接调节输出多样性。
短期落地场景包括交互式写作助手（利用并行生成与局部编辑能力）、代码自动补全（带语法约束的生成）以及广告文案生成（多属性条件控制）。长期愿景是打破自回归模型在文本生成的主导地位，实现与多模态生成范式的统一。

10 结论

连续扩散语言模型通过将离散文本映射到连续空间，成功将扩散模型的并行生成、纠错能力和灵活条件控制引入 NLP 领域。尽管在似然估计上仍落后于自回归模型，其在可控生成和编辑任务上的独特优势已展现出广阔应用前景。随着长文本建模和推理加速技术的成熟，CDLM 有望成为下一代文本生成架构的重要候选方案。