研　究　报　告

离散与连续潜在空间中的扩散语言模型：LLaDA 与 Cola DLM 的对比研究

——基于最大似然估计与层次化信息分解的技术哲学分析

报告编号：〔2026〕第 05-14 号 | 机密等级：公开

摘　要

本研究对两篇具有里程碑意义的扩散语言模型论文——LLaDA（*Large Language Diffusion Models*, 2025）与 Cola DLM（*Continuous Latent Diffusion Language Model*, 2026）——进行了系统深入的对比分析。二者同为非自回归生成范式的代表，却在技术路径上做出了截然不同的选择：LLaDA 坚守离散词元空间，通过掩码扩散（Masked Diffusion）逼近最大似然估计；Cola DLM 则转向连续潜在空间，以层次化信息分解（Hierarchical Information Decomposition）重构文本生成过程。本文从数学形式化基础、建模哲学、架构设计、实验验证及未来影响五个维度展开论述，在忠实还原原文关键公式与数据的同时，力求以通俗语言揭示两种进路背后的深层逻辑。研究发现，二者的分野本质上是关于"语言生成中*什么是最本质的状态空间*"这一本体论问题的不同回答。

关键词：扩散语言模型；LLaDA；Cola DLM；掩码扩散；连续潜在空间；非自回归生成；对比研究

一、问题源起：自回归范式的瓶颈与扩散模型的崛起

大语言模型（LLMs）的成功几乎完全建立在一个核心假设之上：语言可以分解为从左到右的词元链式条件概率。

$$p_\theta(x) = p_\theta(x_1) \cdot \prod_{i=2}^{L} p_\theta(x_i \mid x_1, \ldots, x_{i-1}) \quad \text{（自回归公式）}$$

这一范式——即"下一个词元预测"（next-token prediction）——催生了 GPT、LLaMA、Qwen 等一系列主流大模型。然而，其固有局限也日益凸显：生成顺序被强行绑定为从左到右，导致推理无法并行、反向推理任务（如"B 是 A"）表现

不佳，即所谓的"反转诅咒"（Reversal Curse）。Berglund 等人（2023）的研究表明，若模型在训练数据中见过"A is B"，却无法在推理时正确回答"B is A"。

扩散模型（Diffusion Models）源于热力学非平衡过程的启发，最初在图像生成领域大放异彩（Ho et al., 2020; Song et al., 2020），随后被引入文本领域。其核心思想是：定义一个前向过程逐渐将数据破坏为噪声，再学习一个逆向过程从噪声恢复数据。问题在于——文本是离散的符号序列，而扩散模型天然作用于连续空间。围绕这一矛盾，研究者开辟出两条路径：一条将扩散过程定义在离散词元空间（以 LLaDA 为代表），另一条则将文本先映射到连续潜在空间再施加扩散（以 Cola DLM 为代表）。这两条路径看似目标一致——超越自回归——实则触及了语言建模的根本问题：**语言的内在状态空间究竟是什么？**

二、核心思想：掩码观察恢复 vs. 潜在先验输运

2.1 LLaDA：离散掩码扩散中的最大似然逼近

LLaDA 由中国人民大学高瓴人工智能学院与蚂蚁集团联合提出，发表于 NeurIPS 2025。其技术路线的核心可以概括为一句话：**用掩码预测替代自回归链式分解**。

【外行解释】想象你在玩一个填字游戏：一篇文章中随机位置被涂黑，你的任务是根据可见的上下文同时预测所有被遮住的词。LLaDA 做的事情本质上就是这个——不过它不只是一次性填空，而是从"全部涂黑"开始，逐步"擦掉涂黑处"（即从 t=1 到 t=0 的逆向过程），每一步都重新预测所有剩余被遮住的位置。

数学上，LLaDA 定义了一个**前向数据掩码过程**：对于序列 x_0，时间 $t \in [0,1]$ 时，每个词元被独立地以概率 t 替换为掩码标记 [M]。由此得到条件分布：

$$q_{t|0}(x_t^i \mid x_0^i) = \{\, 1-t, \text{若 } x_t^i = x_0^i; \quad t, \text{若 } x_t^i = [M] \,\}$$

训练目标是极小化仅在掩码词元上计算的交叉熵损失：

$$L(\theta) = -E_{t,x_0,x_t}\left[(1/t) \cdot \sum_{i=1}^{L} 1[x_t^i=M] \log p_\theta(x_0^i|x_t) \right]$$

这一损失函数被证明是模型分布负对数似然的上界（即 $-E[\log p_\theta(x_0)] \leqslant L(\theta)$），使得 LLaDA 成为一个有原则的（principled）生成模型。这一性质极为重要——它意味着优化该损失等价于在变分意义上逼近最大似然估计。尤为关键的是，它与 BERT 的固定掩码比例不同——LLaDA 的掩码比例 t 在 [0,1] 之间均匀采样，这一"微小但关键"的差异赋予了它生成模型的理论地位。

【专业细度】LLaDA 的损失函数在形式上与 MaskGIT 相似，但缺少了 1/t 项。原始论文特别指出（原文 Sec. 2.1）：*"MaskGIT adopts a heuristic training*

*objective, which misses the 1/t term compared to Eq. (3), and lacks a theoretical link to maximum likelihood."*正是最大似然估计这一理论基础，激励了研究团队将离散扩散模型扩展到 8B 规模。

一个更深入的理论洞察是：LLaDA 的掩码扩散模型与任意阶自回归模型（Any-order AR, AO-ARM）之间存在深刻的等价关系。原始论文的附录 A.2 证明，AO-ARM 的期望负对数似然损失（对均匀分布的所有排列求期望）与 LLaDA 的掩码扩散损失在数学上等价。这意味着 LLaDA 实际上隐式地学习了所有可能的生成顺序的加权集成——这解释了为什么它能够在正向和逆向任务上表现一致，而标准的从左到右自回归模型则天然地偏向从左到右的顺序。

在预训练的具体实现上，LLaDA 8B 采用 32 层 Transformer，隐层维度 4096，32 个注意力头，FFN 维度 12288。与 LLaMA3 8B 相比，FFN 维度稍小（12288 vs. 14336），这是因为它使用了普通多头注意力而非分组查询注意力（GQA），导致注意力层参数更多。词表大小为 126,464（LLaMA3 为 128,000）。训练数据中约 11% 为中文、61% 为英文、28% 为代码，经过 PDF 文本提取、去重、有害内容过滤等多道清洗工序。

2.2 Cola DLM：连续潜在空间中的先验输运

Cola DLM 由字节跳动 Seed 团队联合香港大学、澳大利亚国立大学、北京大学提出，发表于 2026 年 5 月。其核心洞察与 LLaDA 截然不同：**扩散过程不应被用于词元级别的观测恢复，而应被用于潜在先验的输运**。

【外行解释】把写文章比作盖房子。自回归模型是一块砖一块砖地从左到右砌墙——每放一块砖都要回头看看之前砌的。LLaDA 的填字游戏式方法相当于同时修正所有墙砖的位置，但每个位置最终还得是具体的砖。而 Cola DLM 则先画设计图（潜在语义变量），再让施工队照着设计图建房子（条件解码）。画设计图和砌墙是两套不同的人马——设计图是连续空间中的抽象表征，砌墙是离散词元的具象实现。

数学上，Cola DLM 是一个层次化潜在变量模型。设 x 为文本，$z_0 \in \mathbb{R}^d$ 为其连续潜在变量。生成模型分解为：

$$p(x, z_0) = p_\theta(x \mid z_0) \cdot p_\psi(z_0), \qquad p(x) = \int p_\theta(x \mid z_0)\, p_\psi(z_0)\, dz_0$$

训练目标为证据下界（ELBO）：

$$\log p(x) \geqslant E_{q_\phi(z_0|x)}[\, \log p_\theta(x \mid z_0) + \log p_\psi(z_0) - \log q_\phi(z_0 \mid x)\,] =: L_{ELBO}(x)$$

展开期望后可得三个分量：

$$E[L_{ELBO}(x)] = E[\log p_\theta(x|z_0)] - I_q(X; Z_0) - KL(\bar{q}_\phi(z_0) \,/\!/\, p_\psi(z_0))$$

这三个分量分别对应：**条件重构**（decoder 能力）、**信息压缩**（z_0 的信息量）和**先验匹配**（学习到的先验与聚合后验的 KL 散度）。这一分解是 Cola DLM 的理论基石——它将文本建模清晰地区分为全局语义组织（潜在先验）和局部文本实现（条件解码）两个层次。

2.3 统一马尔可夫路径视角

Cola DLM 论文提出了一个极为精妙的统一分析框架。设 $\tau = (S_t)_{t\in T}$ 为某个状态空间上的随机过程，则生成模型可统一写作：

$$p_\Theta(x) = \int e_\Theta(x \mid \tau)\, P_\Theta(d\tau)$$

不同模型的差异在于**状态空间**和**路径角色**——而非仅仅是生成顺序。

表1　统一马尔可夫路径视角下的文本模型比较（Cola DLM 原文 Table 1 整理）

模型	状态空间	路径角色	显式潜在变量
自回归 (AR)	前缀词元	直接生成路径	✗
LLaDA	离散掩码序列	离散观测恢复路径	✗
Plaid	连续词元对齐表示	连续观测恢复路径	✗
Cola DLM	压缩潜在序列	**先验输运路径**	√

【哲学高度】这一框架的意义远超技术对比。它揭示了一个根本性问题：当我们用马尔可夫路径来建模文本时，路径的本质是什么？自回归路径是被观测文本本身（词元前缀），LLaDA 和 Plaid 的路径是文本的损坏-恢复轨迹（尽管状态空间不同），而 Cola DLM 的路径既不是文本也不是文本的加噪版本，而是**潜在语义的输运过程**——文本生成只是输运完成后的"副产物"。

三、模型架构与训练机制的系统比较

3.1 整体架构对比

表2　LLaDA 与 Cola DLM 架构详细对比

维度	LLaDA（2502.09992v3）	Cola DLM（2605.06548v1）
全称	Large Language Diffusion with mAsking	Continuous Latent Diffusion Language Model
来源机构	中国人民大学 & 蚂蚁集团	字节跳动 Seed & 港大/澳国立/北大
发表时间	NeurIPS 2025（2025.10 更新）	arXiv 2026.05

模型规模	8B 参数	~2B 参数（500M VAE + 1.8B DiT）
训练数据	2.3T tokens	与 OLMo 2 同源数据，最大长度 512
计算资源	0.13M H800 GPU hours	最高约 2000 EFLOPs
状态空间	**离散词元空间**	**连续潜在空间 $\mathbb{R}^d$**
扩散目标	**词元观测恢复**	**潜在先验输运**
基础架构	Transformer（无因果掩码）	Text VAE + Block-causal DiT + Decoder
训练损失	掩码交叉熵（似然上界）	ELBO（重构 + KL + Flow Matching + 参考正则）
先验学习	隐式（由掩码预测器参数化）	显式（Flow Matching 拟合聚合后验）
SFT	√（4.5M 对话对）	—
RL 对齐	—	—

3.2 训练机制差异

LLaDA 的训练遵循经典的自回归语言模型预训练-SFT 流程，但核心算法完全不同。对于每个训练样本 x_0：

（1）从 [0,1] 中均匀采样 t；

（2）每个词元独立以概率 t 被掩码，得到 x_t；

（3）将 x_t 送入无因果掩码的 Transformer，同时预测所有被掩码的词元；

（4）计算仅针对掩码位置的交叉熵损失（含 1/t 权重）。

训练过程中采用了 Warmup-Stable-Decay 学习率调度器：前 2000 步线性升至 4×10^{-4}，稳定至 1.2T tokens 后降至 1×10^{-4} 再稳定 0.8T，最后 0.3T 线性退火至 1×10^{-5}。优化器为 AdamW，权重衰减 0.1，批量大小 1280。

【关键设计选择深度解析】

Cola DLM 的 Text VAE 采用了严格的因果编码器和因果解码器，这是防止信息泄露的关键设计。与传统的 VAE 不同，它的压缩率仅为 1:1（patch size=1 时一个 token 对应一个潜在变量），并不对序列长度进行压缩——但这并非疏忽，而是有意为之：Cola DLM 的"压缩"不是指序列长度的减少，而是指**语义维度**的压缩：潜在维度 d=16 远小于词表大小约 12.8 万，信息瓶颈迫使模型学习文本的全局语义结构。

另一个精妙的设计是 BERT 掩码损失的作用。它防止 VAE 编码器在语义上坍塌（semantic collapse）——即编码器学会忽略语义信息而解码器单纯记忆表层文本。掩码损失强制编码器保留足够的语义信息以恢复被遮住的词元，从而确保潜在空间中编码的不是微不足道的表面特征，而是真正的语义内容。从某种意义上说，这个设计将 BERT 式"理解"纳入了 VAE 式"生成"的框架中。

Cola DLM 的训练分为两个阶段：

阶段一：Text VAE 预训练。学习文本与潜在变量之间的稳定映射。编码器 $q_\phi(z_0|x)$ 将文本映射到潜在空间，解码器 $p_\theta(x|z_0)$ 从中重建文本。损失函数包含重构损失、KL 正则化和 BERT 风格的掩码损失（防止编码器语义坍塌）。编码器和解码器均为严格因果（causal）结构，防止信息泄露。

阶段二：联合训练。Text VAE 与 Text DiT 同时优化，联合目标为：

$$L_{stage2} = \lambda_{VAE}[-log\ p_\theta(x|z_0) + \beta\ KL(q_\phi \mathbin{/\!/} p_{base}) + \lambda_{mask} L_{mask}] + \lambda_{fm} L_{FM} + \lambda_{ref}\ E[KL(q_\phi \mathbin{/\!/} q_\phi^{ref})]$$

其中 Flow Matching 损失 L_{FM} 用于学习块级条件先验，参考正则项 $E[KL(q_\phi \mathbin{/\!/} q_\phi^{ref})]$ 抑制联合训练中的潜在漂移。关键设计是 Block-causal DiT：块内双向注意力 × 块间因果依赖——这一结构与 Eq. (3.3) 的块级先验分解严格一致。

3.3 推理机制对比

LLaDA 的推理：从完全掩码的序列开始，通过多步去掩码逐步生成文本。支持三种采样策略：纯扩散采样、自回归采样和块扩散采样。采用低置信度重掩码策略（low-confidence remasking）——每一步将预测概率最低的 st 个词元重新掩码，类似 LLM 采样中的退火技巧。默认使用均匀时间步长，生成长度作为可调超参数。

Cola DLM 的推理：先将前缀编码为干净的潜在条件，然后在潜在空间中逐块生成（每个块通过 Flow Matching ODE 从噪声输运至语义），最后通过条件解码器生成文本。推理过程无需在每个去噪步骤中处理完整序列的离散词元——DiT 在潜在空间中以 8–10 步即可完成大部分有效去噪（原文 Fig. 9a），对应理想条件下比自回归解码减少 1.6–2.0× 的顺序生成深度。CFG 的调整在推理阶段至关重要（原文 Fig. 9b）：任务平均分在 CFG 从 0 增至 3–6 时快速上升，在约 7 处达到最优，之后过大值（>10）导致急剧下降。

值得注意的是，LLaDA 也支持 CFG（原文 Appendix B.3），且 CFG 在选定的六个基准（ARC-C、HellaSwag、TruthfulQA、WinoGrande、PIQA、GPQA）上持续改善性能。例如，ARC-C 从 45.9 提升至 47.9，HellaSwag 从 70.5 提升至 72.5。但由于要与自回归模型进行公平比较，主实验结果中未使用 CFG。两种模型对 CFG 的兼容性表明，扩散语言模型天然支持这种灵活的推理时控制机制。

四、实验验证与定量比较

4.1 LLaDA 的关键实验发现

可扩展性。LLaDA 在 1B 和 8B 两个规模上的可扩展性实验表明，其性能趋势与自回归基线高度一致，在 MMLU 和 GSM8K 等任务上甚至展现出更强的可扩展性。论文在 10^{20}~10^{23} FLOPs 范围内验证了这一结论（原文 Fig. 3），这是此前 Nie et al. (2024) 在 10^{18}~10^{20} 范围内结论的显著扩展。

基准测试结果。预训练的 LLaDA 8B Base 使用 2.3T tokens 训练，在 MMLU（65.9 vs. 65.4）、GSM8K（70.3 vs. 48.7）、Math（31.4 vs. 16.0）、CMMLU（69.9 vs. 50.7）上全面超越 LLaMA3 8B Base（原文 Tab. 1）。值得注意的是，GSM8K 上 70.3 vs. 48.7 的显著优势说明 LLaDA 在数学推理任务上具有特殊优势。

反转诅咒的破解。在诗歌补全任务中（496 对著名中文诗句），LLaDA-8B Instruct 在正向生成上得分为 51.8（略低于 GPT-4o 的 82.7 和 Qwen2.5 的 75.9），但在逆向生成上达到 45.6，远超 GPT-4o 的 34.3 和 Qwen2.5 的 38.0（原文 Tab. 4）。这一结果具有重大理论意义——它实证性地证明，"反转诅咒"并非语言模型的固有宿命，而是自回归范式特定偏置的副产品。论文指出，LLaDA 在训练中词元被无偏地对待，没有"从前到后"的归纳偏置，因此自然地实现了正向和反向性能的平衡。

SFT 之后的多轮对话能力。经过 SFT 后（4.5M 对话对，3 个 epoch），LLaDA 展现出令人印象深刻的指令遵循能力。Tab. 3 展示了一个四轮对话案例：用户先用英文询问诗句，再要求翻译为中文、德文，最后要求创作一首以 C 开头的五行诗。LLaDA 全部正确完成，且在德语翻译中保持了罗伯特·弗罗斯特《未选择之路》原文的意境。这是首次有非自回归模型展现出如此连贯的多轮对话能力。Tab. 2 表明，即使仅使用 SFT（无 RL），LLaDA 8B Instruct 在 ARC-C（88.5 vs. 82.4）和 GPQA（33.3 vs. 31.9）上超越 LLaMA3 8B Instruct（使用了 SFT+RL），这进一步证实了扩散路径的潜力。

采样灵活性与效率权衡。LLaDA 原生支持三种采样模式（Appendix B.4）：自回归采样（逐词元）、块扩散采样（组内扩散 + 组间自回归）和纯扩散采样（同时预测所有词元）。实验表明（原文 Tab. 7 & Tab. 8），纯扩散采样在 Base 模型 5 个任务上均取得最佳综合表现（如 BBH 49.7、GSM8K 70.3、HumanEval 35.4），而块扩散 LLaDA（block size=32）在 Instruct 模型的 GSM8K（77.5）和 Math（42.2）上甚至优于纯扩散（69.4 和 31.9）。这种灵活性是 LLaDA 的独特优势——用户可以根据任务需求在质量和速度之间自由权衡。

4.2 Cola DLM 的关键实验发现

全局语义结构的存在性。Cola DLM 通过"最优 timeshift"实验为潜在空间中的全局语义结构提供了定量证据。实验表明，随着潜在维度 d 从 16 增加到 64 再到 128，最优 timeshift 从约 1.0 漂移至 1.7 再到 2.3（原文 Fig. 2）。这一系统性漂移与"潜在表征是纯局部且完全可分"的零假设相矛盾，证明潜在空间中存在跨维度共享的语义结构。

潜在空间设计的最优策略。实验比较了五种策略：固定 VAE、联合训练（lr 同比例）、联合训练（lr 0.01×）、随机初始化联合训练、间隔式训练。结果表明（原文 Fig. 3）：最好的策略既非保持潜在空间固定，也非从零联合训练，而是在稳定初始化基础上与 DiT 共同演进。随机初始化的潜在空间在 d=16 时显著坍塌，增大至 d=128 虽部分缓解但结构仍不可比。

DiT 块大小与噪声调度。块大小 16 综合表现最优（原文 Fig. 6），过大的块（64/128）显著降低语义交互质量。噪声调度方面，logit-normal 分布 loc=1.0 在 30K 和 40K 两个检查点均取得最佳平均性能（原文 Fig. 8）。

整体缩放表现。在 8 个基准上的统一评估（原文 Fig. 10）显示：Cola DLM 在任务平均上实现了最强的整体缩放趋势。在平均任务得分上，Cola DLM 在整个计算预算范围内稳步提升，最终取得了最佳性能。AR 在小预算下保持竞争力，LLaDA 在早期也有明显提升，但 Cola DLM 的曲线在高计算预算区间上升最为持续。尤其在依赖全局语义组织的推理密集型任务（MMLU、RACE、Story Cloze、OBQA）上表现突出——MMLU 上持续上升并在中后期超过 AR；RACE 和 Story Cloze 上在中高计算预算区间持续保持最优。在生成式任务（LAMBADA 和 SQuAD）上，Cola DLM 也展现出令人鼓舞的缩放行为：LAMBADA 上稳步提升并接近 AR，SQuAD 上最终超越 AR 并逼近 LLaDA。需注意，由于采用统一的生成式评估协议（非标准的似然分类评估），多选基准的绝对分数较低，但相对缩放趋势公平可信。原始论文特别指出（原文 Sec. 4.5），将潜在维度从 16 提升至 128 已证实可提高语义容量，因此当前结果是"保守配置"下的表现。

4.3 似然评估与生成质量的结构性错位

Cola DLM 论文的一个深刻发现是：对于连续潜在语言模型，**生成质量与似然导向的困惑度之间不存在必然对应关系**（原文 Sec. 5.1）。在局部潜在几何分析中（原文 Fig. 11 & Tab. 4），解码器探针成功率和后验命中率持续较高，但先验命中率变化剧烈。这意味着问题的根源不是解码器失效，而是后验邻域附近的先验错位。

例如，对于同一目标词"at"，固定 VAE logSNR=1.0 时似然困惑度从 1.15×10^6 改善至 641.57，但生成词却从正确的"on"退化为"in"。这表明：**更低的困惑度并不必然带来更好的生成质量**——前者要求围绕真实后验的精确局部密度校准，而后者只要求先验质量覆盖解码器有效区域。

【哲学高度】这引出了一个元层次的认识论问题：当我们用不同的评估标准来评判模型时，我们究竟在评判什么？对于自回归模型，训练目标与评估指标天然对齐（都是词元级别的最大似然）。对于层次化潜在模型，模型在优化的实际上是一个"不同"的东西——它学习的是潜在语义的组织方式，而非离散分布的精确拟合。因此，**缩放行为而非单一困惑度值，才是更能反映此类模型真实潜力的指标**。

五、技术哲学视野下的深层对比

5.1 对"文本本质"的本体论分歧

在潜在稳健性方面，Cola DLM 做了系统性的评估（原文 Sec. 5.4 & Fig. 13）。VAE 在 t=0 时重建准确率达 0.9998，表明潜在-文本映射高度保真且未坍塌。更重要的是，重建准确率在低噪声区间保持极高水准：当扩散时间步 t=250（即注入 25% 噪声强度）时，准确率仍约 0.92。在前 40% 的噪声区间内呈现优雅的渐进退化（graceful degradation）模式，而非突变式坍塌。这意味着潜在空间对扰动具有高度稳健性，足以作为后续先验建模的语义接口。

压缩实验（原文 Sec. 5.3 & Tab. 6）则揭示了更深层的洞察。当 patch size=2（每两个 token 压缩为一个潜在变量）时，整体性能低于 p=1，但按序列长度奇偶性拆分后发现：序列长度为偶数（Mod0）时，p=2 的平均得分（18.12）甚至超过了 p=1（17.31）——这说明压缩本身不会损害性能，当前局限来自对不可整除序列边界的不稳定处理。这一发现与 Cola DLM 的核心原则一致：潜在空间的价值不在于无损保存词元级信息，而在于提供更低信息率的全局语义组织。

LLaDA 和 Cola DLM 之间的根本差异，不是"离散 vs. 连续"这样一个技术层面的选择，而是关于**语言的本体论地位**的根本分歧。

LLaDA 的隐含立场：文本就是词元序列。生成文本就是选择正确的词元序列。扩散模型的作用是找到一种比从左到右更好的方式来采样词元序列。因此，LLaDA 的所有操作都在离散词元空间中进行：前向过程将词元替换为 [M] 标记，逆向过程预测词元类别，损失函数衡量词元预测的交叉熵。

Cola DLM 的隐含立场：文本包含两个不同层次的信息——**全局语义**（"这段话大概在说什么"）和**局部实现**（"用哪些具体的词来表达"）。前者本质上是连续的、低维的、可压缩的；后者是离散的、高维的、需要精细解码的。因此，一个"诚实"的生成模型应当显式区分这两个层次，而不是用一个单一的自回归链将它们混为一谈。

后者可以从信息论的角度严谨表述。Cola DLM 论文提出了"三条控线"（three governing curves）的判断准则：

$$\textit{Cola DLM 具有优势} \Leftrightarrow D(R) \text{ 在低 } R \text{ 时已经很小} \quad \wedge \quad E(M_{ColaDLM}) \text{ 持续下降} \quad \wedge \quad G_{infer} \text{ 可控}$$

其中 D(R) 是表征率失真函数。这意味着 Cola DLM 的成功并非天然保证——它取决于数据是否真的具有"低维全局语义 + 高维局部实现"的层次结构。如果数据不存在这种结构（比如所有信息都不可压缩地分布在所有词元中），则层次化潜在模型反而会因信息瓶颈而受损。

5.2 对"学习"的认识论差异

LLaDA 的认识论立场是**经验主义**的：学习就是最小化预测误差，通过足够的算力和数据，模型自然涌现出各种能力。这也是当前主流大模型的基本哲学——可扩展性假说（Scaling Hypothesis）。

Cola DLM 则更接近**理性主义**的立场：学习不仅是拟合数据，更是发现数据的内在结构。通过强制引入一个信息瓶颈（潜在变量 z_0），模型被迫去发现文本中真正"重要"的全局语义结构，而不是仅仅记忆词元之间的统计相关性。这反映在 ELBO 分解中的 $I_q(X; Z_0)$ 项——它在显式地限制潜在变量承载的信息量。

正如 Cola DLM 论文的 Afterword（Sec. 8）所深刻阐述的：学习是一个"模型-环境交互系统"，由三个因素共同决定——（1）模型如何表征和吸收信息（状态空间）；（2）环境如何定义改进方向（评估指标）；（3）环境本身的真实结构（世界是否具有跨模态的联合规律）。这已经超越了单纯的技术讨论，进入认知科学和系统论的核心领域。

LLaDA 论文的结论则更加务实而有力（原文 Sec. 5）：*"Our findings show the promise of diffusion models for language modeling at scale and challenge the common assumption that these essential capabilities are inherently tied to ARMs."*它并不试图构建新的认识论框架，而是用实验事实证明：扩散模型也能做到自回归模型能做的事——这本身就是一种深刻的挑战。

5.3 对多模态扩展的视野差异

LLaDA 对多模态的展望相对保守（原文 Sec. 5）：*"LLaDA's ability to process multi-modal data remains unexplored."*这由其技术路线决定——在离散词元空间中的掩码扩散并未天然提供与图像、视频等连续模态的接口。

Cola DLM 则从架构上天然支持多模态扩展。其核心公式——模态特定的 VAE 编码/解码 + 共享的块因果先验——可以直接扩展到图像模态（原文 Sec. 5.5 & Fig. 14）。论文给出了文本续写、图像条件文本生成和文本到图像生成的初步定性结果，并提出了一个形式化的统一框架：设 z_0^{text} 和 z_0^{img} 为文本和图像的潜在变量，联合生成过程可写作：

$$p(x_{text}, x_{img}, \tilde{z}_0) = p_\theta(x_{text}, x_{img} \mid \tilde{z}_0) \cdot p_\psi(\tilde{z}_0), \qquad \tilde{z}_0 = (z_0^{text}, z_0^{img})$$

这一扩展路径连接了离散文本与连续多模态系统，指向了更具雄心的"统一世界模型"目标。原文 Sec. 8.3 从模型-环境交互的系统论视角对此进行了阐释：如果真实世界中的观察、转移和反馈通常是由共享的潜在状态联合决定的，那么将不同模态映射到共享的连续潜在空间中，就不仅是工程上的便利，而是对世界真实结构的更准确建模。

原始论文以形式化语言表述了这一观点。设环境 $E = (\Omega, O, A, T, F, G)$，其中 Ω 为环境状态空间，O 为观察空间，A 为动作空间，T 为状态转移机制，F 为反馈生成机制，G 为将反馈转化为优化信号的规则。当观察是多模态的且真实环境的反馈机制依赖于跨模态联合状态而非各模态独立时，统一建模就不仅是效率问题，而是学习正确世界模型的必要条件。

六、综合评述与展望

核心结论。LLaDA 和 Cola DLM 代表了非自回归文本生成的两种根本不同的范式：

LLaDA 证明了**离散掩码扩散可以在大模型尺度上与自回归全面竞争**，其最大贡献在于挑战了"核心 LLM 能力天然依赖于自回归模型"的普遍假设。它以 8B 参数、2.3T tokens 的训练预算，达到了与 LLaMA3 8B 相当的性能，并从根本上解决了反转诅咒问题。

Cola DLM 则提出了一种**更彻底的替代方案**——不仅丢弃从左到右的生成顺序，更重新定义了文本生成的目标本身：不是恢复被损坏的词元，而是输运潜在语义的分布。它在 ~2B 参数规模上展现了令人鼓舞的缩放行为，并为其理论框架提供了实证支撑。

6.1 各自的局限

LLaDA 的局限（原文 Sec. 5）：（1）生成长度是超参数——虽然模型对此不敏感，但自适应长度会更好；（2）受限于计算资源，与自回归模型的直接对比仅限制在 10^{23} FLOPs 以下；（3）没有 KV cache 等系统级优化；（4）尚未进行 RL 对齐；（5）缩放范围（2.3T tokens）仍远小于顶级自回归模型（15T+）。

Cola DLM 的局限（原文 Sec. 6）：（1）实验规模仍然较小（~2B），需要更大参数和计算预算下的验证；（2）潜在空间的稳定性依赖于 VAE logSNR、DiT 块大小、噪声调度的精细联合校准；（3）潜在压缩方案在不被序列边界整除的情况下性能不稳定的问题未完全解决；（4）似然评估与生成质量之间的结构性错位使得模型评估缺乏统一的量化标准。

6.2 未来方向

两种路径的未来发展呈现出有趣的互补性。LLaDA 可以探索的方向包括：引入 RL 对齐、扩展到更大规模（10B+）、结合专门的位置编码和注意力机制、以及更高效的采样算法。Cola DLM 的前沿则包括：更强的潜在模块（如 AE、RAE）、更灵活的先验学习方法、以及向多模态领域的系统性扩展。

【综合判断】从现有证据来看，LLaDA 的技术成熟度更高（已到 8B 规模，经 SFT 后可进行多轮对话），更接近实用。Cola DLM 的理论框架更深刻（统一马尔可夫路径视角、率失真分析、评估错位理论），但距离实际部署还有距离。二者并非竞争关系，而是揭示了扩散语言模型设计中一个根本性的"分裂"：**扩散过程应当作用于哪个层次——词元还是语义？**这个问题的答案将深刻影响大模型未来的发展方向。

6.3 对语言模型研究范式的启示

这两篇论文共同指向了一个可能正在发生的范式转变。自 2018 年 GPT 和 BERT 以来，语言模型的研究几乎完全被"自回归"和"去噪自编码"两种预训练目标所主导。LLaDA 和 Cola DLM 开启了第三条道路：**扩散路径上的非自回归生成**。

从更广阔的视角看，Cola DLM 论文的 Afterword（Sec. 8）提出了一个值得深思的框架：将学习形式化为一个模型-环境交互系统 Mθ 在环境 E 中的优化过程。在这个框架下，"自回归 vs. 扩散"之争被重新诠释为关于**表征空间、目标函数和环境结构**三个层面的联合选择。自回归语言模型在三个维度上做出了自洽的选择——离散词元表征、直接似然最大化、纯符号环境——而 Cola DLM 则同时在三个层面上改变了假设：引入连续潜在层次、以 ELBO 代替直接似然、为离散文本提供通往连续多模态环境的接口。

如果从更宏大的科学史视角来看，这场争论实际上是 20 世纪认知科学中"符号主义 vs. 联结主义"之争在深度学习时代的回响。符号主义认为智能可以归结为离散符号的操作（自回归 + 离散词元 = 符号串的生成），联结主义则认为智能源于连续空间中分布的、亚符号的相互作用（潜在扩散 + 连续语义 = 分布的输运）。两种范式各有其深刻洞察，但也都各有局限。LLaDA 和 Cola DLM 的研究表明，语言建模的未来可能不在于固守某一端，而在于理解何时以及如何在离散符号和连续语义之间建立有效的映射和协同。

6.4 未竟之问与开放问题

尽管两篇论文都取得了令人瞩目的成果，但它们也各自留下了悬而未决的关键问题，这些问题的答案将决定扩散语言模型未来的走向。

缩放效率之争。Nie et al. (2024) 曾指出，掩码扩散模型需要 16 倍于自回归模型的计算量才能达到相同的似然值。LLaDA 论文将缩放范围从 10^{18}~10^{20} FLOPs 扩展到了 10^{20}~10^{23} FLOPs，并在下游任务上展示了竞争力。但这是否意味着在更大规模上（10^{24} FLOPs 以上）仍然成立？Cola DLM 提供了另一个答案：如果模型不是

在词元级别恢复观测，而是在潜在空间中输运语义先验，那么"似然"这个指标本身可能就不适用了。但这需要更大规模实验的验证。

评估范式之困。Cola DLM 论文揭示的似然-生成质量错位提出了一个深层问题：对于非自回归生成模型，什么才是正确的评估指标？如果困惑度不能反映真实能力，生成式评估又难以标准化和自动化，那么我们如何公平地比较不同模型？这不仅仅是技术问题，更是计量学（metrology）的问题——当我们试图衡量"智能"时，我们需要确保尺子本身是准确的。

架构统一之可能。有趣的是，Cola DLM 论文在 Table 1 中将 LLaDA 列为"离散观测恢复路径"的代表，并将其与自己的"先验输运路径"区分开来。但 LLaDA 论文的 Transformer 设计——没有因果掩码、可以同时看到所有输入——实际上与 Cola DLM 中 DiT 的块内双向注意力有相通之处。是否存在一种统一的架构，可以同时支持词元级的掩码预测和潜在级的流匹配？两篇论文都没有回答这个问题，但它无疑是未来研究的重要方向。

【数据支撑总结】

两篇论文的关键数据对比如下：LLaDA 以 8B 参数、2.3T tokens、0.13M H800 GPU hours 的训练投入，在 MMLU (65.9 vs. 65.4)、GSM8K (70.3 vs. 48.7)、CMMLU (69.9 vs. 50.7) 等核心基准上超越 LLaMA3 8B Base；在反转诗歌补全任务上以 45.6 分大幅超越 GPT-4o 的 34.3 分。Cola DLM 以 ~2B 总参数、在 2000 EFLOPs 预算内，通过层次化潜在建模在任务平均上取得最优缩放趋势，并在推理密集型任务（MMLU、RACE、Story Cloze、OBQA）上持续保持优势，8–10 步推理即可完成大部分有效去噪。两条路径共同证明：非自回归的扩散语言模型正在从理论可能性走向实际可行性。

参考文献

[1] Nie, S., Zhu, F., You, Z., et al. *Large Language Diffusion Models*. NeurIPS 2025. arXiv:2502.09992v3.

[2] Guo, H., Zhao, Q., Zhao, Y., et al. *Continuous Latent Diffusion Language Model*. arXiv:2605.06548v1, 2026.

[3] Berglund, L., et al. *The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"*. arXiv:2309.12288, 2023.

[4] Austin, J., et al. *Structured Denoising Diffusion Models in Discrete State-Spaces*. NeurIPS 2021.

[5] Shi, J., et al. *Simplified and Generalized Masked Diffusion for Discrete Data*. arXiv:2406.04329, 2024.

[6] Ho, J., et al. *Denoising Diffusion Probabilistic Models*. NeurIPS 2020.

[7] Kingma, D.P. & Welling, M. *Auto-Encoding Variational Bayes*. ICLR 2014.

[8] Dubey, A., et al. *The Llama 3 Herd of Models*. arXiv:2407.21783, 2024.

[9] Touvron, H., et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv:2307.09288, 2023.

[10] Chang, H., et al. *MaskGIT: Masked Generative Image Transformer*. CVPR 2022.

研究单位：AI 技术前沿分析课题组

完成日期：2026 年 5 月 14 日

（本报告共约 8000 字，基于对原始论文的忠实解读与分析）