扩散语言模型调查

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen

Abstract—扩散语言模型（DLMs）正迅速成为一种强大且有前途的替代当前主流自回归（AR）范式的方法。通过迭代去噪过程并行生成标记，DLMs 在减少推理延迟和捕获双向上下文方面具有内在优势，从而能够对生成过程进行精细控制。在实现几倍速度提升的同时，最近的进展已使 DLMs 显示出与其自回归对手相当的性能，使其成为各种自然语言处理任务的一个引人注目的选择。尽管其日益普及，DLMs 呈现出既有挑战又有机遇，这需要对其原理、技术和局限性进行详细和系统的理解。在本调查中，我们对当前的 DLM 领域提供了一个全面的概述。我们追溯了它的演变及其与其他范式，如自回归和掩码语言模型的关系，并涵盖了基础原理和最新的模型。我们的工作提供了一个最新的、全面的分类法，并对当前技术进行了深入分析，从预训练策略到先进的后训练方法。此调查的另一个贡献是对 DLM 推理策略和优化的深入审查，包括对解码并行性、缓存机制和生成质量的改进。我们还强调了 DLMs 多模态扩展的最新方法，并描绘了它们在各种实际场景中的应用。此外，我们的讨论还涉及了 DLMs 的局限性和挑战，包括效率、长序列处理和基础设施需求，同时勾勒未来的研究方向以维持这一快速发展的领域的进步。项目的 GitHub 地址为 https://github.com/VILA-Lab/Awesome-DLMs 。

Index Terms—Diffusion Language Model, Large Language Model, Diffusion Model, Diffusion Large Language Model, Language Modeling, Multimodal Language Model

1 介绍

最近通用人工智能（AGI）的进步主要由自回归大型语言模型（LLMs）[1]–[7] 和图像及视频生成的扩散模型 [8]–[12] 的出现所驱动。这些模型在理解和生成各种模态方面表现出非凡的能力，实现了以前难以想象的性能水平。这些模型的规模空前巨大，从大量的参数数量、庞大的数据集、训练时的巨大努力、到推理过程中显著的计算需求，推动了人工智能达到新的高度，为这些模型提供了广泛的通用知识以及对语言和真实世界的深刻理解。

GPT 系列的兴起，尤其是 ChatGPT 的公开发布，推动了自回归（AR）语言模型在自然语言处理领域占据主导地位。通过使用因果注意力和教师强制来训练预测下一个标记，AR 模型能够有效地扩展到大型数据集和模型规模。AR 模型通过逐个标记的顺序方式生成文本，擅长支持从简单问答到复杂推理和创意写作的广泛任务。然而，这种顺序性质对推理速度造成了主要瓶颈。自回归生成过程一次生成一个标记，天然限制了并行性，并显著限制了计算效率和吞吐量。

扩散模型是另一种极具前景的生成性范式。它们通过一个去噪的过程，从逐渐加噪声的版本中恢复数据，并通过逐步逆转这种随机损坏的步骤来生成新的样本。在复杂数据分布建模方面表现出色，扩散模型在图像和视频合成中已达到最先进的结果 [13] 。在扩散建模方面的学术突破 [14]–[17] 为训练和推断建立了坚实的理论基础。同时，像 Stable Diffusion [8], [10], [11] 、Imagen [9] 和 Sora [12] 这样的大规模实用模型展示了扩散范式的卓越可扩展性和泛化能力，能够从简单的文本提示（通常只需要几个词语）中生成高保真、艺术级的图像和视频。除了在复杂数据分布建模上的强大能力外，扩散模型在并行性方面提供了固有优势。通过一个迭代的去噪过程，它们可以同时生成多个标记或整个序列，这可能导致更优的推断吞吐量和更好地利用现代并行计算硬件。尽管仍面临一些挑战，尤其是在离散数据建模和处理动态序列长度方面，扩散语言模型（DLMs）作为应对生成质量与速度之间权衡的一种有力替代方案已经出现。

• *Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen are with VILA Lab, Mohamed bin Zayed University of Artificial Intelligence. Mingda Chen is also with Department of Automation, Tsinghua University. E-mail: { Tianyi.Li, Bowei.Guo, Zhiqiang.Shen } @mbzuai.ac.ae, cmd22@mails.tsinghua.edu.cn*

为了使扩散适应于离散语言数据，已经提出了几种关键的方法。在早期，DLMs 的发展主要受扩散模型在图像合成等连续领域成功的驱动。连续 DLMs 将标记映射到嵌入中，并在连续空间中执行去噪，如先驱工作 Diffusion-LM [18] 和 SED [19] 所示。而离散 DLMs 则直接在标记空间中定义扩散过程。D3PM [20] 等早期工作引入了带有吸收状态的结构化转移矩阵，允许标记级别的破损和迭代去噪。后续的工作如 DiffusionBERT [21] 整合了预训练的掩码语言模型（例如，BERT）以提升去噪质量，并提出了量身定制的噪声计划（例如，纺锤计划）以更好地使标记损坏与标记频率对齐。这些早期模型证明了在非自回归文本生成中应用迭代去噪的可行性，提供了可控性和并行性，尽管它们的性能仍然落后于强大的自回归基线。随着核心挑战在 DLMs 中逐步得到解决，并且这种范式趋于成熟，更大规模的 DLMs 已被开发。通过从自回归模型开始初始化，像 Dream [22] 和 DiffuLLaMA [23] 等 7B 级模型已经表明 DLMs 可以从现有模型中有效调整，同时实现具有竞争力的性能。LLaDA-8B [24] 进一步展示了从头训练 DLMs 的潜力，取得了与类似规模的 LLaMA3-8B 模型相当的性能。多模态 DLMs，也称为扩散多模态大型语言模型（dMLLMs），在建模文本和图像等混合数据方面也显示出潜力。基于开源 DLMs，像 LLaDA-V [25] 、Dimple [26] 和 MMaDA [27] 等模型将跨模态推理和生成整合到扩散框架中。同时，工业努力也显示出对 DLMs 的越来越大的兴趣。Mercury 系列 [28] 和 Gemini Diffusion [29] 在实现每秒数千个标记的推理速度的同时报告了强大的性能。这些发展突显了 DLMs 日益增长的实用性和商业潜力。我们在 Fig. 1 中提供了 DLMs 发展的时间线，从代表性模型到最近的进展 [30]–[36] ，随后在 Fig. 2 中可视化了 DLM 的趋势。

扩散语言模型在训练和推理中也呈现出独特的挑战和机遇。预训练通常遵循与自回归语言模型或图像扩散模型相似的策略 [22], [26], [27] 。为了加速训练并重用之前的训练成果，许多 DLMs 是从预训练好的自回归模型权重初始化的 [22], [23] 。DLM 中的监督微调（SFT）也类似于自回归模型：提供干净的提示数据，模型学习生成目标完成。强化学习（RL）也在训练后应用于 DLM，以提高复杂任务的表现。已经提出了 GRPO [37] 算法的变体，如 diffu-GRPO [38] 和 UniGRPO [27] ，以增强 DLM 大规模的推理能力和对齐。在推理过程中，已经开发了各种策略和优化方法来充分利用 DLM 的能力。连续 DLM 可以利用 ODE/SDE 求解器或其他少步生成

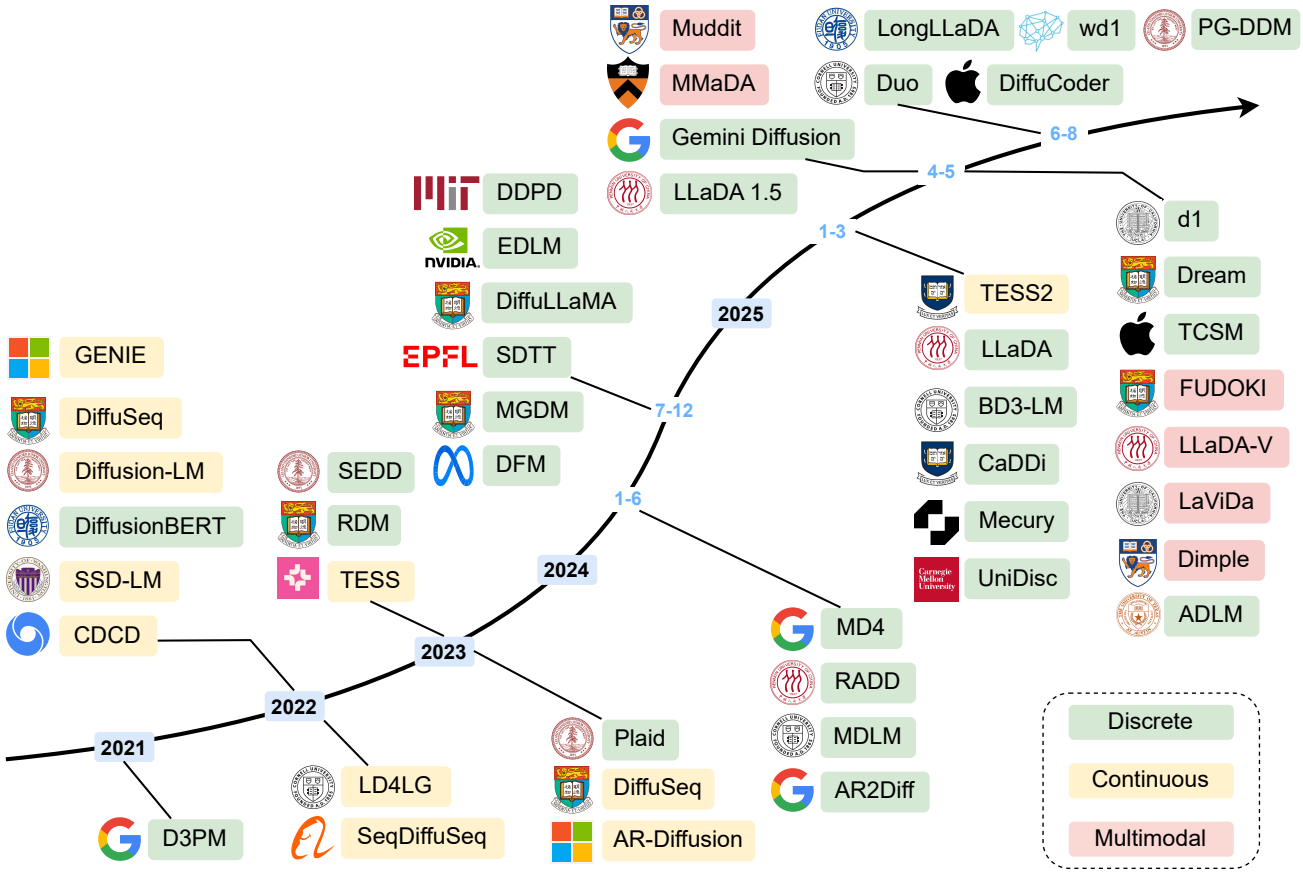

Fig. 1. 扩散语言模型的发展时间线。该图突出了 DLM 发展的关键里程碑，被分为三组：连续 DLM、离散 DLM 和近期的多模态 DLM。我们观察到，尽管早期研究主要集中于连续 DLM，离散 DLM 近年来越来越受欢迎。

技术来加速迭代去噪过程 [39] 。由于离散 DLM 在并行生成中面临更多挑战，已提出专用的并行解码策略 [26], [40], [41] ，以允许在单步中接受多个标记并克服并行的障碍。去遮盖和重新遮盖策略 [24], [42] 通过选择性地揭示低置信度标记进一步改善生成质量，而缓存技术 [43], [44] 可以显著减少计算并增强两种范式的推理速度。

与自回归模型相比，扩散语言模型被广泛认为具有以下几个明显优势：

- 并行生成：DLMs 可以通过迭代去噪过程并行生成多个标记，相较于自回归模型显著提高推理速度和吞吐量。
- 双向上下文：DLMs（双向语言模型）自然地整合了双向上下文，能够实现更加细致入微的语言理解和生成。它们还能够生成更丰富的上下文嵌入，这对跨模态生成任务非常有利。这也使得对生成过程进行细粒度的控制成为可能。
- 迭代细化：迭代去噪过程允许 DLMs 在多个步骤中更新他们的感知。通过尽早接受高置信度的 token 并将低置信度区域保留为遮掩，遮掩的 DLMs 可以逐步改善不确定区域，通常会产生更连贯和更高质量的文本生成。
- 可控性：DLMs 可以根据特定的标记位置或结构进行条件设置，使其非常适合填充和结构化生成等任务。此外，指导技术（例如，无分类器指导）可以更好地控制风格和语义相关性。
- 跨模态的统一建模：通过应用一个共享的去噪建模框架，DLM 自然支持统一的文本和视觉生成任务。这使得它们在需要在单一模型中同时进行生成和理解的多模态应用中尤其有前景。

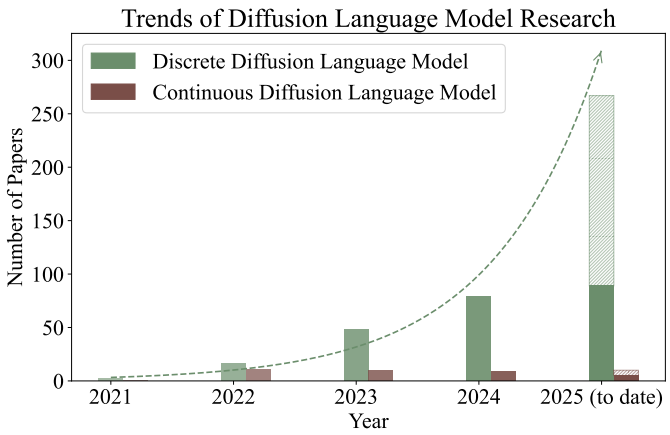

Fig. 2. 扩散语言模型论文的趋势。对于离散 DLM，统计数据来自引用 D3PM [20] 的论文，并进一步选择那些标题或摘要中包含关键词“language”的论文。对于连续 DLM，统计数据基于与本文相关的资料库中记录的相关研究数量。结果反映了该领域研究兴趣的增长。统计数据仅供参考。

尽管近期深度学习模型（DLMs）的流行度有所上升，但仍缺乏系统覆盖整个 DLM 生态系统的综合调查。我们的调

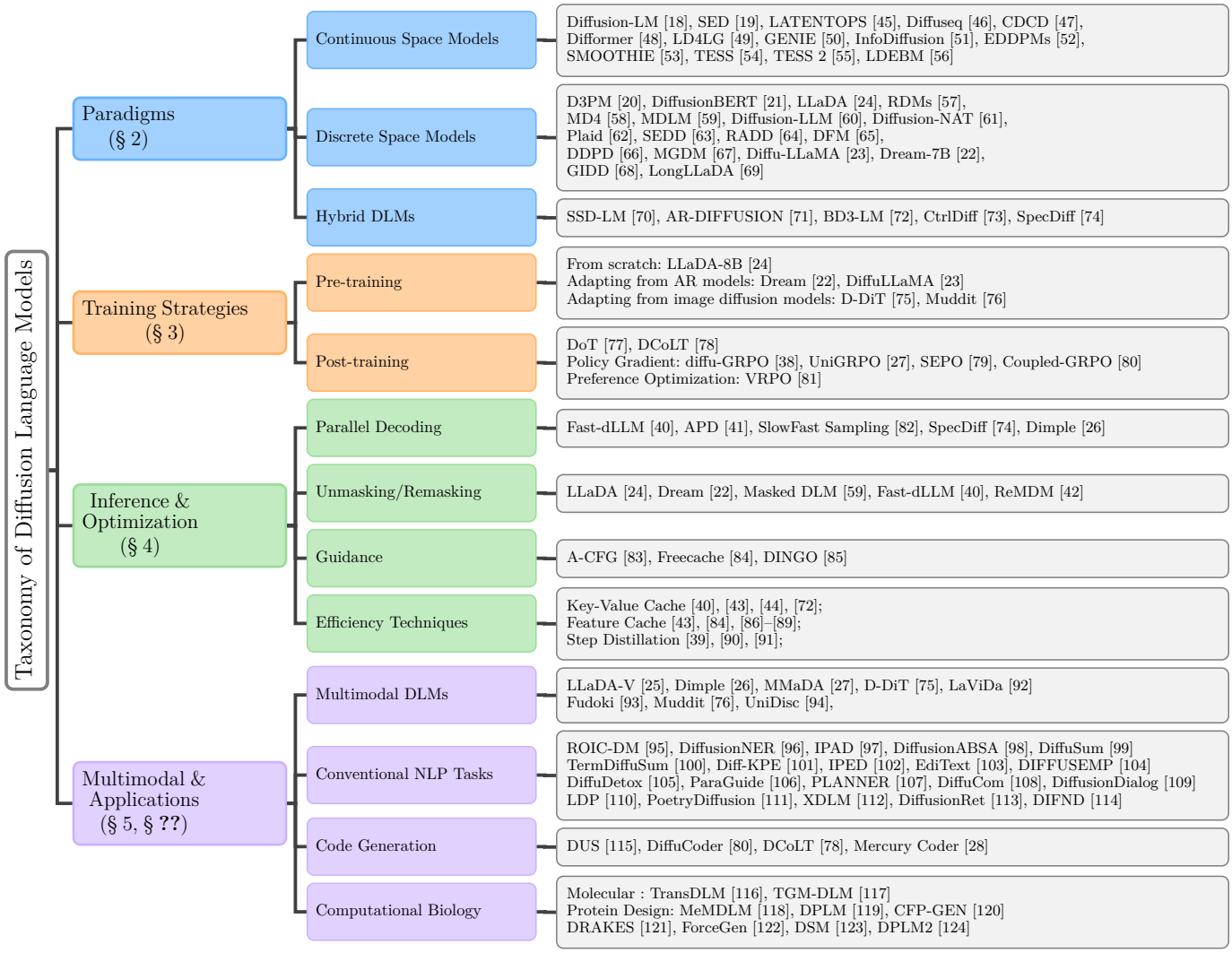

Fig. 3. 扩散语言模型的分类法，涵盖基础、训练和推理策略，以及关键应用。章节编号（§）对应于本调查中的各个部分。

查结构如下：第 2 节全面概述了现代语言建模范式，包括自回归、掩蔽和基于扩散的方法。第 3 节深入探讨了扩散语言模型的训练方法，涵盖了预训练和后续微调技术，如 SFT 和 RL 对齐。第 4 节详细介绍了各种推理策略和优化，重点是针对连续和离散空间模型的技术。第 5 节探讨了扩散模型在多模态上下文中的扩展，调查了如 LLaDA-V [25] 、MMaDA [27] 和 Dimple [26] 等最先进的模型和架构。第 ?? 节展示并可视化了 DLMs 的性能比较。第 ?? 节展示了 DLMs 在从文本和代码生成到计算生物学等任务中的多样应用。第 7 节强调了扩散语言模型的挑战和局限，包括效率、推理、代理能力和基础设施问题，还概述了未来研究的有前景方向。为了提供综合概述，图 3 中展示了 DLMs 的分类。

2 扩散语言模型的范式

扩散语言模型作为一种强大的非自回归范式出现，能够在生成质量和推理并行性之间取得平衡。受非平衡热力学原理的启发 [125] ，DLMs 学会逆转逐渐的加噪过程。这种迭代的改进方法允许对整个序列进行并行生成，为 AR 模型的推理瓶颈提供了一种潜在的解决方案。DLMs 大致可根据扩散过程运行的空间分为连续或离散。此外，还有结合自回归和扩散的混合模型，它们通过不同的形式旨在利用这两种范式的互补优势。我们在表 1 中展示了多个作品的模型信息，并在图 4 中提供了不同范式的比较。

2.1 现代语言建模的初步研究

语言建模领域经历了几个不同的范式，每个范式都具有独特的架构选择、训练目标和相关的权衡。在本小节中，我们简要概述了近期基于 Transformer 的大规模范式，重点介绍其核心原则、数学公式和具有代表性的模型。由于我们在此专注于现代的大规模设计，因此不包括早期的方法。此综述旨在建立概念基础，以便理解扩散语言模型作为一种新颖且有前景的替代方案的出现，该方案解决了先前方法的关键限制。

2.1.1 掩码语言模型

被 BERT 推广的掩码语言模型（MLMs），代表了一种基础范式，该范式使用基于 transformer 的仅编码器架构缩放预训练语言模型。MLMs 在概念上简单但在经验上强大，通过预测输入序列中随机掩码的标记来学习双向上下文表示，利用了前后的上下文。这种方法遵循去噪自编码框架，其中输入标记的子集被掩码，模型被训练来重建它们：此处，x 表示输入序列，$\mathcal{M}$ 是掩码位置的集合，$x_{\setminus\mathcal{M}}$ 代表可见（未掩码）上下文。BERT 还引入了一个下句预测（NSP）目标来建模句子间关系：其中 (A, B) 是一对文本片段，$y \in \{0, 1\}$ 表示 B 是否在原始文本中跟随 A 。

BERT 在语言理解任务中的有效性，例如情感分析、命名实体识别和问答，激发了众多改进版本的出现。例如，RoBERTa [126] 去除了 NSP 目标，并采用了更激进的训练策略，而 ALBERT [127] 引入参数共享和矩阵分解以提高效率。DeBERTa [128] 通过解耦注意力和改进的掩码标记预测解码机制进一步增强了上下文编码。

尽管 MLM 在理解任务方面具有优势，但它们本质上并不是为生成任务而设计的，生成文本需要专门的微调策略或解码方案，在没有重大架构修改的情况下，它们不适合开放式生成。

以 GPT 系列 [1], [2], [129], [130] 和 Transformer-XL [131] 为例，经过后续大规模语言模型（LLMs）[3]–[5], [132] 的进一步发展，自回归语言模型已成为现代生成式 AI 的基石，其特点是单向的、从左到右的标记生成过程。与双向模型不同，自回归语言模型通过将文本序列的联合概率分解为条件概率的乘积来进行建模：

$$P(x) = \prod_{i=1}^{n} P_\theta(x_i \mid x_1, x_2, \ldots, x_{i-1}) \tag{1}$$

给定一个标记序列 $X = (x_1, x_2, \ldots, x_n)$，训练目标是在这种分解下最大化序列的对数似然性：

$$\mathcal{L}_{\text{AR}} = \mathbb{E}_{X\sim\mathcal{D}}\left[-\sum_{i=1}^{n} \log P_\theta(x_i \mid x_1, \ldots, x_{i-1})\right] \tag{2}$$

这通常使用仅解码器的 Transformer 架构来实现，在训练期间采用因果注意力遮蔽和教师强制，确保每个标记的预测仅依赖于之前的标记，同时使损失的并行计算成为可能。

顺序生成的形式既是一个优势，也是一个限制。一方面，它与文本生成任务对齐，便于直接采样，自然适用于各种应用。另一方面，它对推理速度构成了一个根本性的瓶颈，因为标记生成本质上是顺序的，无法并行化。这种生成质量与延迟之间的权衡已成为推进自回归（AR）模型的核心挑战。除了标准的下一个标记预测（NTP），最近的研究探索了多标记预测（MTP）[133], [134]，以通过每步生成多个标记来加速推理。这些努力在概念上与去噪语言模型（DLMs）中采用的并行解码策略有相似之处。

2.1.2 其他范式

序列到序列模型。序列到序列 (Seq2Seq) 模型 [135]，作为一种早期但强大的范式，基于编码器-解码器架构构建，并作为用于条件文本生成任务的多功能框架，例如机器翻译和摘要。现代模型如 T5 [136] 和 BART [137] 是杰出的例子。

在这个架构中，编码器处理源序列以生成中间表示，解码器则利用该中间表示生成目标序列，通常是以自回归的方式。虽然标准的 Seq2Seq 解码器是自回归的，但该框架本身具有高度灵活性。许多深度语言模型，例如 DiffuSeq [46] 和 SeqDiffuSeq [138]，通过用非自回归扩散解码器替代自回归解码器来适应这一架构，利用编码器的强条件能力来指导生成过程中的去噪过程。

排列语言模型。排列语言模型（PLM），以 XLNet [139] 为例，提供了一种在生成框架中结合双向上下文的替代方法。PLM 被训练用来预测序列中的标记，但不是按照固定的从左到右的顺序，而是以随机的、排列的顺序。目标是最大化因子顺序的所有可能排列的期望对数似然：

$$\mathcal{L}_{\text{PLM}} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_T}\left[-\sum_{t=1}^{N} \log P_\theta(x_{z_t} | \boldsymbol{x}_{\boldsymbol{z}_{<t}})\right] \tag{3}$$

其中 $\mathcal{Z}_T$ 表示长度为 T 的序列的所有可能排列的集合，$\boldsymbol{z}_t$ 和 $\boldsymbol{z}_{<t}$ 代表给定排列 $\boldsymbol{z} \in \mathcal{Z}_T$ 的第 t 个和第一个 $t-1$ 元素。该公式使得模型能够为每个标记捕获双向上下文，结合了双向上下文（如 MLM）与连贯的自回归生成过程的优势。这与 DLM 形成对比，DLM 通过并行迭代细化过程实现双向性。

2.2 连续扩散语言模型

连续空间的 DLM 通过首先将离散标记映射到连续嵌入空间来建模语言。然后扩散过程在这个连续空间中建模数据分布 [18], [19] 。通常，扩散模型通过学习逆转一个预定义的损坏过程来定义生成过程，该过程逐渐将数据转化为噪声。该过程包括一个前向（加噪声）过程和一个逆向（去噪声）过程。前向过程通过一个固定的马尔可夫链在 T 个时间步中逐渐将数据样本 $\boldsymbol{x}_0$ 转化为噪声：

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \tag{4}$$

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \mu_t(\boldsymbol{x}_{t-1}), \Sigma_t), \tag{5}$$

其中 μ_t 和 Σ_t 定义了噪声时间表。在许多实现中，如 DDPM [14] 和 Rectified Flow [17]，每个时间步处的边际分布都有一个封闭形式的表达式：

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + b_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}), \tag{6}$$

其中 α_t 和 b_t 是时间 t 的确定性函数。

逆过程学习逆转损坏，从噪声 $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ 开始，逐渐去噪以恢复接近 $\boldsymbol{x}_0$ 的样本。该过程由神经网络 $f_\theta(\boldsymbol{x}_t, t)$ 参数化，通常实现为 Transformer，它预测与前向过程相关的目标量 $\boldsymbol{z}$（例如，干净数据、噪声或速度）。一个常见的训练目标的形式是：

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{z}}\left[\|f_\theta(\boldsymbol{x}_t, t) - \boldsymbol{z}\|^2\right], \tag{7}$$

，其中 $\boldsymbol{x}_t$ 是通过给定 $\boldsymbol{x}_0$ 的前向过程采样得到的，而 $\boldsymbol{z}$ 是从 $\boldsymbol{x}_0$ 和 t 派生的相应的回归目标。

训练完成后，通过从学习到的反向过程采样来进行生成，从噪声 $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ 开始。在每个时间步 $t = T, T-1, \ldots, 1$，模型定义一个条件分布 $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$，旨在逼近真实的反向转换 $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$。从这些学习到的条件分布中逐步采样会生成逐渐噪声减少的潜在状态，直到恢复原始数据的估计 $\boldsymbol{x}_0$。在生成去噪嵌入 $\hat{\boldsymbol{x}}_0$ 后，一个舍入步骤将其映射回离散的标记。这通常通过在嵌入空间中的最近邻搜索或使用解码头来完成。

Diffusion-LM [18] 首次在嵌入空间中引入了扩散过程，以创建非自回归的语言生成模型。通过使用类似于图像扩散模型中的分类器引导机制，它实现了高度可控的文本生成和填充。LDEBM [56] 在变分学习框架中提出了一种潜在空间能量模型与扩散模型的新型共生，以解决基于能量的先验的学习问题，重点在于可解释的文本建模。LATENTOPS [45] 提出了一种在紧凑潜在空间中操作的高效可组合文本操作框架。它引入了一种基于常微分方程 (ODE) 的高效采样器，以产生由任意插件控制算子引导的潜在向量，然后将其解码为所需的文本。之后，Diffuseq [46] 提出了一种无分类器的序列到序列任务的扩散语言模型（DLM），其在前向过程中仅破坏目标序列嵌入，以实现强大且多样的条件文本生成。自条件嵌入扩散（SED） [19] 框架直接在固定的、连续的标记嵌入空间中进行扩散。通过结合自我条件机制，它在条件和无条件文本生成中均表现出强大的性能，媲美标准自回归模型。CDCD [47] 将连续扩散应用于分类数据，通过将标记嵌入到连续空间中。它提出了得分插值，这种独特的方法允许模型通过交叉熵损失进行训练，以及时间变形，这种自适应策略可高效地在训练期间调度噪声水平。为了应对嵌入空间中的优化挑战，Difformer [48] 引入了一种锚点损失以防止嵌入崩溃，并提出了一种噪声重新缩放框架以缓解模型退化。LD4LG [49] 利用

TABLE 1
扩散语言模型、配置及其设计选择的总结。

Model	Parameters	Diffusion type	Noise schedule	Task	Training data
D3PM [20]	70M	Discrete	Mutual information	Language	65B tokens
Diffusion-LM [18]	100M & 300M	Continuous	Square-root	Language	–
Diffuseq [46]	91M	Continuous	Square-root	Language	565K sentence pairs
SSD-LM [70]	400M	Continuous	Cosine	Language	123B tokens
DiffusionBERT [21]	110M	Discrete	Spindle	Language	16B tokens
CDCD [47]	1.3B	Continuous	–	Language	315B tokens
LD4LG [49]	188M	Continuous	Cosine	Language	5.2M sentence pairs
SeqDiffuSeq [138]	65M & 110M	Continuous	Adaptive	Language	45B tokens
TESS [54]	125M & 355M	Continuous	Linear	Language	–
MDLM [59]	110M	Discrete	Log-linear	Language	622B tokens
DFM [65]	1.7B	Discrete	Linear & Cubic	Language & Code	2.5T tokens
TESS-2 [55]	7B	Continuous	Log-linear	Language	360B tokens
LLaDA [24]	1B & 8B	Discrete	Linear	Language & Code	2.3T tokens
Mecury [28]	–	Discrete	Log-linear	Code	Trillions tokens
LLaDA-1.5 [81]	8B	Discrete	Linear	Language	2.3T tokens
MMaDA [27]	8B	Discrete	Log-linear	Language	900B image-text tokens
Dream [22]	7B	Discrete	Log-linear	Language & Code	580B tokens
LLaDA-V [25]	8.4B	Discrete	Linear	Multimodal	3M image-text samples
LaViDa [92]	8.4B	Discrete	Convex	Multimodal	1.6M image-text samples
Dimple [26]	7B	Discrete	Log-linear	Multimodal	0.8B tokens
LongLLaDA [69]	8B	Discrete	Log-linear	Language & Code	2.3T tokens
DiffuCoder [80]	7B	Discrete	Log-linear	Code	130B tokens

预训练语言模型作为强大的自动编码器，以创建一个紧凑的潜在空间，随后在其中训练连续扩散模型以实现高质量文本生成。GENIE [50] 提出了一种用于扩散语言模型的大规模预训练框架，引入了一种新颖的连续段落去噪目标，通过重建损坏的文本段落来有效地从大型语料库中学习。InfoDiffusion [51] 引入了一种信息熵感知的噪声计划，以引导模型朝着更类似人类的“关键信息优先”过程，该过程优先生成核心内容。EDDPMs [52] 通过使用参数化的编码器-解码器，将生成、重建和表示统一到一个框架中，从而推广扩散过程，实现所有组件的稳定联合训练。SMOOTHIE [53] 提出了一种新颖的扩散过程，它基于语义相似性逐步平滑标记嵌入，结合了连续潜在空间和离散标记处理的优点。

连续扩散过程也可以在对数空间而不是嵌入空间中进行。TESS [54] 引入了一个完全非自回归的框架，该框架在标记的 k-对数简单形表示上扩散，并采用了针对这一设置的全新的自我条件机制。进一步拓展，TESS 2 [55] 通过扩展预训练的大型自回归模型到通用扩散语言模型，实现了方法的规模化，在扩散特定的预训练方案和指令微调中实现，赋予强大的指令跟随能力。

2.3 离散扩散语言模型

离散空间的 DLMs 直接在标记词汇上定义了扩散过程，避免了在扩散过程中需要一个连续的嵌入空间。D3PM [20] 首先通过引入在离散标记上的结构化扩散过程来说明这一点。前向过程通过在每一步应用一个转移矩阵 $\boldsymbol{Q}_t$ 来破坏序列。该矩阵定义了词汇中一个标记转移到任何其他标记的概率。给定初始状态 $\boldsymbol{x}_0$ 的状态 $\boldsymbol{x}_t$ 的概率通过一个分类分布给出：

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathrm{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_0\bar{\boldsymbol{Q}}_t), \quad \text{where} \quad \bar{\boldsymbol{Q}}_t = \prod_{i=1}^{t} \boldsymbol{Q}_i$$

对于 $\boldsymbol{Q}_t$ 一个常见的选择是一个吸收状态转移，其中每个标记都有一种保持不变或转移到特殊 ‘[MASK]‘ 标记的概率。反向过程学习逆转这些转移，预测给定破坏序列的原始标记的概率分布。随着时间的推移，掩码 DLMs 已成为离散扩散语言模型的现代且高度有效的进化，构成了几个最近的大规模努力的基础 [23], [24] 。我们以 LLaDA [24] ，这种类型中最具代表性的模型为例。受到早期关于重新参数化和简化训练目标的工作的启发 [57], [58], [64] ，LLaDA 从零开始训练，使用仅在掩码标记上计算的交叉熵损失：

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t,x_0,x_t}\left[\frac{1}{t}\sum_{i=1}^{L} 1[x_t^i = \mathrm{M}] \log p_\theta(x_0^i|x_t)\right], \quad (8)$$

其中 x_0 是从训练语料库中采样的，t 从 $[0,1]$ 中均匀采样，x_t 是通过前向过程破坏 x_0 而获得的。指标函数 $1[\cdot]$ 确保损失只施加在被掩码的位置。在推理过程中，生成过程以所需长度的全掩码序列开始。在每个迭代步骤中，模型采用当前序列（包含生成的标记和 ‘[MASK]‘ 标记的混合体），并预测完整的标记序列。根据模型的预测置信度和噪声计划，将特定数量的最高置信度预测解除掩码并固定，而剩余位置则重新掩码。此精炼过程不断迭代，直到所有的 ‘[MASK]‘ 标记被解决。这种方法巧妙地结合了 MLM 的双向上下文与可控的并行生成过程。特别是，LLaDA-8B 展示了强大的可扩展性和指令遵循能力，其表现可与强大的自回归模型如 LLaMA3-8B 相媲美。这挑战了长期以来自回归模型在大规模语言生成中的主导地位。DiffusionBERT [21] 结合了预训练的 BERT 与离散扩散过程，利用其强大的去噪能力从被屏蔽的状态中学习逆过程。通过一种考虑标记信息的新颖主轴噪声调度进一步增强该模型，与此前的 DLM 相比，在生成质量上取得了显著的改进。另一种方法，重新参数化离散扩散模型（RDMs） [57] ，建立了逆过程的替代公式，从而将训练目标简化为加权交叉熵损失。这使得更灵活和自适应的解码策略成为可能，相较于之前的离散扩散模型获得了显著的性能提升。同样地，MD4 [58] 推导出了一个简单的加权交叉熵损失积分，作为屏蔽扩散模型的连续时间变分目标，为训练 DLM 提供了一个简单而通用的框架。另一种类似的方法是 MDLM [59] ，其引入了一种简化的、Rao-Blackwell 化目标，表现为屏蔽语言建模损失的加权平均。Diffusion-LLM [60] 通过将预训练的掩码语言模型适应扩散范式，以及进行任务特定微调和指令微调，展示了 DLM 的可扩展性，解锁了它们在解决一般语言任务中的多功能性。Diffusion-NAT [61] 将离散扩散模型与 PLM 统一，通过将去噪过程重新表述为非自回归的掩码标记恢复任务，使得 BART 可以充当有效的去噪器。Plaid [62] 是第一个通过最大化数据似然来训练的扩散语言模型，通过规模法则证明其可以在标准基准测试中优于 GPT-2 等自回归模型。为了改

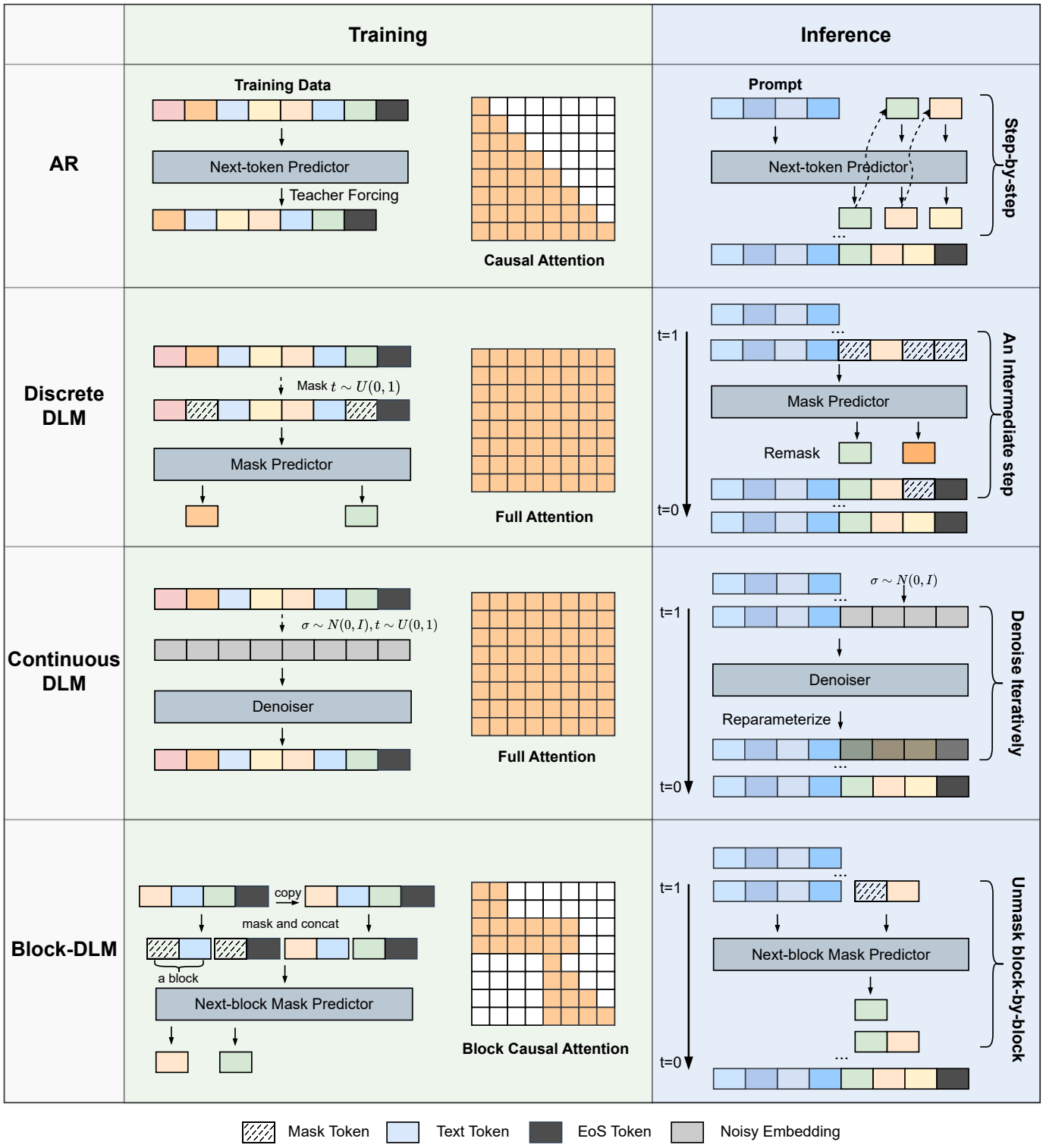

Fig. 4. 扩散语言模型不同范式的训练和推断过程概览，并包括自回归（AR）模型以便于比较。AR 模型使用教师强迫和因果注意力进行训练，而离散和连续的 DLMs 则采用完全双向的注意力机制。以 BD3-LM [72] 为例的分块扩散模型，整合了自回归和扩散策略，并使用专门设计的块因果注意力掩码进行训练。

进训练目标，SEDD [63] 引入了一个分数熵损失，以直接学习数据分布的比率，这作为分数匹配的离散扩展。重参数化吸收离散扩散（RADD） [64] 揭示了吸收扩散中的具体分数可以表示为干净数据的时间不变条件概率，乘以一个解析的、时间依赖的标量。它还正式统一了吸收离散扩散和任意阶自回归模型的训练目标。离散流匹配 (DFM) [65] 引入了一种针对离散数据的新型生成范式，与连续流匹配类似。该方法学习生成概率速度，通过一般概率路径家族中的转变，从源分布到目标分布。通过扩展模型架构，DFM 显著缩小了在各类基准测试中与自回归模型的性能差距。DDPD [66] 提出了一种将生成过程解耦为两个专业模型：规划器和去噪器的框架。在每个步骤中，规划器识别出需要改进的最受损的标记位

置，之后去噪器预测它们的值。为提高复杂推理任务的性能，MGDM [67] 被引入以解决子目标不平衡问题。该方法通过在学习过程中通过标记级重加权机制优先考虑更困难的子目标来增强离散扩散。为应对扩展的挑战，提出了一种持续预训练方法 [23]，以改造现有自回归模型，如 LLaMA，转换为扩散语言模型。最终的模型，命名为 DiffuGPT 和 DiffuLLaMA，与其自回归对等模型竞争，同时获得扩散原生能力，如灵活补全。在此观察基础上，Dream-7B [22] 从 Qwen2.5 7B [140] 初始化，并通过 5800 亿个标记进行了进一步训练，远远超过现有的 DLMs，匹配顶级自回归模型的性能。GIDD [68] 被引入，以克服掩码扩散模型无法修改生成标记的限制。该框架通过将掩码与均匀噪声结合，推广了噪声过程，这解锁了模型自我纠正错误的能力，并提高了样本质量。最近，为了解决长文本能力问题，LongLLaDA [69] 在该领域提供了第一个关于 DLMs 的系统分析。它揭示了 DLMs 可以在直接上下文外推时保持稳定困惑度，并具有更好的检索能力。LongLLaDA 还引入了一种基于 NTK 的 RoPE 无训练外推方法，这显著提升了 DLMs 的外推性能，验证了已建立的外推缩放规律对于 DLMs 仍然有效。

2.4 混合 AR-扩散语言模型

混合自回归-扩散模型旨在平衡非自回归模型的完全并行性和自回归模型的强因果依赖性建模。在混合自回归-扩散建模中，一个显著策略是采用分块的半自回归生成过程。在这一设置中，模型自回归地生成令牌块，而每个块内的令牌使用类似扩散的迭代过程并行生成。早期的尝试如 SSD-LM [70] 通过在单纯形表示上的块状连续扩散过程开创了混合方法，AR-DIFFUSION [71] 则展示了多级扩散过程，并通过根据令牌位置调整时间步来实现半自回归。最近的代表模型 BD3-LM [72] 在离散模型上更进一步推进了这一方向，表现出与纯自回归和扩散模型相比的强劲性能。CtrlDiff [73] 通过引入动态块预测技术来提高块级效率和控制，从而改进了这一范式。

在这些模型中，生成过程通常由两个嵌套的循环组成。在外层循环中，使用自回归的方式生成令牌块，每个块都依赖于先前生成的块。在每个块内，内层循环通过类似扩散的迭代去噪过程并行地逐个生成令牌。在 BD3-LM 中，训练目标被形式化为：

$$\mathcal{L}_{\mathrm{BD}}(\boldsymbol{x},\theta) := -\sum_{b=1}^{B} \mathbb{E}_{t\sim[0,1]} \mathbb{E}_{q} \frac{1}{t} \log p_\theta(\boldsymbol{x}^b | \boldsymbol{x}_t^b, \boldsymbol{x}^{<b}) \tag{9}$$

这种混合策略使模型能够通过自回归捕获跨块的长距离依赖，同时通过并行扩散加速每个块内的生成。该设计还支持灵活的输出长度和在 AR 模型中广泛使用的 KV 缓存 [72]。

值得注意的是，最近的掩码扩散语言模型 [24], [27] 也采用了类似的半自回归基于块的解码策略，这可以视为混合 AR-扩散建模的实例。

除了在序列级别结合自回归和扩散的块级方法外，混合也可以发生在架构级别，其中神经网络的某些部分（通常是编码器）将整个序列扩散到一个中间表示，然后由自回归解码器生成最终序列 [141]。LADIDA [142] 是一种略有不同的方法，在文档级别扩散，但通过自回归解码器解码句子。SpecDiff [74] 提出了一种协作推测解码框架，其中轻量级扩散模型起草候选输出，然后由大型自回归模型验证并最终确定输出。

3 DLMs：预训练和后训练

3.1 预训练和监督微调

DLM 的预训练过程主要遵循与自回归语言模型（用于离散 DLM）或图像扩散模型（用于连续 DLM）相似的程序，设计空间相对较少。本节简要总结了现有的 DLM 预训练方法，旨在弥合 DLM 与 AR 模型之间的方法学差距。

为了加速训练，特别是对于大规模模型，通常的做法是从预训练的自回归语言模型或图像扩散模型中初始化 DLMs。DiffuGPT 和 DiffuLLaMA 尝试用参数从 127M 到 7B 的开源 LLMs 初始化掩码 DLMs，发现 DLMs 可以高效地从自回归模型适配，显著减少训练时间和成本，同时实现与其自回归对等体相当或甚至更优的性能。基于这一见解，Dream-7B 从 Qwen 2.5 7B 初始化，据报道在各种基准测试中优于 LLaDA-8B 和 LLaMA3-8B。而一些多模态 DLMs 则从预训练的图像扩散模型中初始化。D-DiT 和 Muddit 分别从预训练的 SD3 和 Meissonic 的 MM-DiT 骨干中初始化。尽管这些模型最初并不是为文本生成设计的，但它们的潜在表示包含内在的语言对齐知识，可以有效地促进语言建模的训练，同时保持较强的视觉生成能力。

在 DLM 中的监督微调通常与 AR 模型类似。对于像 LLaDA [24] 这样的掩码 DLM，提示令牌不被掩码，而响应令牌被选择性地掩码，从而使模型能够以一种与预训练兼容的方式学习条件响应生成。在连续 DLM 中，SFT 也可以通过仅破坏响应段来进行，如 TESS2 [55] 所示。

尽管与 AR 训练范式总体相似，但由于采用了基于扩散的公式，DLM 面临几个独特的挑战。一个主要问题在于掩码 DLM 损失计算的效率。在典型的掩码 DLM 训练中，如果均匀采样时间步长，只有大约 50 % 的标记（平均）参与损失计算。这降低了数据的利用率，可能导致次优梯度，特别是如果关键回答标记被排除在损失之外。为了解决这个问题，LaViDa [92] 提出了一个补充的掩码策略：每个训练样本都使用两个不相交的掩码模式进行复制，确保所有标记至少一次被包含在损失计算中。此外，由于如 [143] 所示的训练-推理差异，模型在训练期间的表现明显优于推理时。作者提出了一个两步的扩散过程和一种改进的调度技术来缓解这一问题。

3.2 后训练推理能力

随着它们在语言任务上的表现提升，DLM 中对推理能力的探索变得越来越流行。通常，推理能力是通过在推理数据集上进行微调来获得的。对于 DLM 来说，这提出了一个独特而艰巨的挑战。传统的连锁式思维（CoT）方法基于 AR 模型的顺序特性来逐步推理，但 DLM 却是并行生成 tokens。AR 领域中最成功的训练后技巧，尤其是基于强化学习（RL）和策略梯度方法的那些，都是建立在有效计算生成序列的对数概率的能力之上的。由于 AR 模型的可因式分解和顺序特性，这种计算是简单的。然而在 DLM 中，生成是一个迭代的、非顺序的过程，对数似然是不可计算的，这为将成熟的 RL 算法套件应用于 AR 模型带来了重大的技术障碍。直观地，我们将这些工作分为三个主要方向，这也构成了本小节的结构：(1) 并行化推理链，即在 AR 模型中将 CoT 适配到 DLM 的并行生成。(2) 适配策略梯度方法，即在 DLM 中引入流行算法如 GRPO 的变体。(3) 适配偏好优化方法如 DPO 到 DLM。

3.2.1 DoT 和 DCoLT：并行化推理链

其中一个引领在深度学习模型中引发复杂推理的早期研究是 Diffusion-of-Thought（DoT） [77]，该研究将流行的链式推理范式适应于扩散框架。与自回归模型依序生成推理步骤不同，DoT 将其制定为中间思维，在整个扩散去噪过程中并行完善。这种方法通过对预训练的深度学习模型进行微调来实现，例如 Plaid [62] 和 SEDD [63]，在包含问题及其对应的逐步推理的数据集上训练。为了增强模型从其自身错误中恢复的能力，DoT 引入了定制的训练技术，如定计划采样和耦合采样，这些技术在训练过程中让模型接触到自身生成的错误，从而提高其自我纠正能力。这种后期训练的方法使得

TABLE 2
对当前 DLMs 推理能力的训练后方法进行简要总结，重点介绍它们的算法类型、主要目标、关键技术创新和适用模型类型。值得注意的是，这些方法大多数基于策略梯度，并且是为离散 DLMs 设计的。

Method	Algorithm Type	Core Goal	Key Technical Innovation	Model Type
DoT [77]	Non-RL Fine-tuning	Enable parallel Chain-of-Thought reasoning	Converts serial CoT into a parallel diffusion process; training-time self-correction	Continuous/ Discrete
DCoLT [78]	Outcome-based RL	Enable non-linear latent reasoning	Lateral thought; outcome-based RL; Unmask Policy Module	Continuous/ Discrete
SEPO [79]	Policy Gradient Framework(PPO/GRPO)	Finetune discrete DLMs with non-differentiable rewards	Low-variance gradient estimator via score entropy & importance sampling	Discrete
diffu-GRPO [38]	Policy Gradient (GRPO)	Introduce policy gradient method to DLMs	Efficient one-step log-probability estimator for applying GRPO to masked DLMs	Discrete
coupled-GRPO [80]	Policy Gradient (GRPO)	Reduce variance and maintain training efficiency	Coupled-sampling with complementary masks	Discrete
UniGRPO [27]	Policy Gradient (GRPO)	Unified reinforcement learning	Structured noising strategy; diversified reward modeling	Multimodal Discrete
VRPO [81]	Preference Optimization (DPO)	Align with human preferences	Sample budget allocation; antithetic sampling	Discrete

较小的深度学习模型能够达到令人印象深刻的推理性能，甚至在某些数学和逻辑推理基准上超越体积显著较大的自回归模型。更近期的方法，Diffusion Chain of Lateral Thought（DCoLT） [78]，引入了一种基于强化学习的独特推理框架，这种框架受到侧向思维认知概念的启发，与传统链式推理方法的逐步垂直思维形成对比。DCoLT 不是监督中间步骤，而是将逆扩散过程中的每一步视为潜在的思维动作，并通过基于结果的强化学习优化整个多步骤去噪轨迹，以最大化最终答案的奖励。当应用于如 LLaDA 这种掩码深度学习模型时，DCoLT 创新地引入了去掩蔽策略模块（UPM），该模块学习揭示标记的最佳顺序作为强化学习动作空间的一部分。这一方法显著提升了深度学习模型的推理能力，使得经过 DCoLT 强化的 LLaDA 模型在 GSM8K 上取得了 +9.8 % 的提升，在 HumanEval 上取得了 +19.5 % 的提升。

3.2.2 将策略梯度方法应用于 *DLMs*

分数熵策略优化（SEPO） [79] 引入 RLHF 至离散 DLMs，提出一个理论基础的框架，用政策梯度方法和不可微奖励微调离散扩散模型。在分数熵框架内运行，SEPO 通过使用重要性采样来推导稳定且低方差的梯度估计，从而调整现代政策梯度方法如 PPO 和 GRPO。这样使得模型的策略可以迭代更新以最大化奖赏函数，使其成为用于条件和无条件生成的通用框架。SEPO 的目标函数定义如下：

$$l^A(\theta) = \mathbb{E}_{x\sim\pi_{\theta_{old}}}\left[\sum_{\substack{y\in\mathcal{X}\\ y\neq x}} w_{x,y}\log s_\theta(x, T-T_0)_y\right] \tag{10}$$

模型参数 θ 经过优化以最大化由 $w_{x,y} = \pi_\theta(y) f(r_{x,y}^{T-T_0})$ 加权的分数熵 s_θ 的期望对数似然。期望是对来自之前策略 $\pi_{\theta_{\text{old}}}$ 的样本 x 进行的。可以选择函数 f 来还原不同的策略梯度变体；例如，截断函数会产生 PPO，而标准化组奖励会产生 GRPO。该公式能够实现稳定和低方差的梯度估计，即使在非可微奖励的情况下，并为细化离散扩散模型提供了一个灵活的目标。通过若干离散生成任务的数值实验展示了 SEPO 的可扩展性和效率，证明策略梯度强化学习可以稳健地应用于离散扩散模型。d1 [38] 提供了一种两阶段的后训练框架，用于结合监督微调（SFT）和新颖的策略梯度算法 diffu-GRPO 的掩码 DLM。为了使 GRPO 适应缺乏分解似然的 DLM，d1 引入了序列对数概率和逐个令牌对数概率估计的新方法。d1 使用简单的平均场分解，通过独立的逐个令牌概率的乘积来近似序列对数概率，而逐个令牌对数概率则通过在每次策略梯度更新时，根据随机掩码提示条件下进行完整掩码完成的单次前向传播来计算。在每个内梯度更新步骤中对提示使用不同的随机掩码作为一种正则化，提高了训练效率和稳定性。完整的 d1 流水线，通过 SFT 随后是 diffu-GRPO，在 LLaDA 模型的数学和计划推理任务上展现了显著的性能提升。MMaDA [27]，一个统一的多模态扩散模型，呈现了三阶段的训练流程。在第一阶段预训练之后，MMaDA 采用混合长思维链微调策略，将来自不同任务的推理轨迹精心调整为统一格式，以在跨模态中对齐推理过程。这促进了第三阶段的编码启动训练，该阶段引入了 UniGRPO，一种为扩散语言模型量身定制的策略梯度强化学习算法。UniGRPO 通过利用结构化的噪声策略克服了像 d1 这样的基础方法的局限性，这种策略通过均匀采样一个掩码比例 $p_i \in [0, 1]$ 而不是屏蔽所有的响应标记。这确保了模型在多步扩散去噪过程中暴露于不同的阶段，从几乎完全掩码到几乎未掩码，这与传统的扩散训练一致，并提升了模型多步去噪能力的利用效率。此外，序列级别的对数似然通过对掩码标记的平均来逼近。DiffuCoder [80] 是一个特别为代码生成开发和分析的 7B 参数的 DLM。这项工作引入了一种名为耦合-GRPO 的 RL 算法，该算法设计为通过利用 DLM 生成过程的独特属性来原生支持扩散。耦合-GRPO 的核心创新在于其用于对数似然估计的耦合采样方案。为了获得更稳健和更低方差的估计，它为训练批次中的每个完成序列构建了成对的、互补的掩码。对于给定的序列，生成两种掩码，以使得每个标记位置在这两个掩码中的其中一个中完全被掩盖。然后通过平均这两个互补前向过程的损失来得出对数概率估计。这确保在训练过程中每个标记都在部分掩码的上下文中被评估，与使用单个随机掩码或全掩码的方法相比，提供更全面的标记覆盖和更稳定的梯度信号。耦合-GRPO 显示出显著提高了 DiffuCoder 在代码生成任务上的性能，同时还促进了更多的并行化和较少自回归的生成模式。

3.2.3 将偏好优化应用于 *DLMs*

LLaDA 1.5 [81] 提出了一个名为方差减少偏好优化（VRPO）的新框架，以适应离散 DLM 的偏好优化方法。该研究识别出在离散 DLM 上应用直接偏好优化（DPO）具有挑战性，因为用于近似对数似然的证据下界（ELBO）方差较大。VRPO 通过引入两个关键的无偏方差减少技术来解决这一问题：（1）通过采样更多扩散时间步长而不是每个时间步长的多个掩码版本来最优分配蒙特卡洛采样预算，即 $n_t = n$ 和 $n_{y_t} = 1$

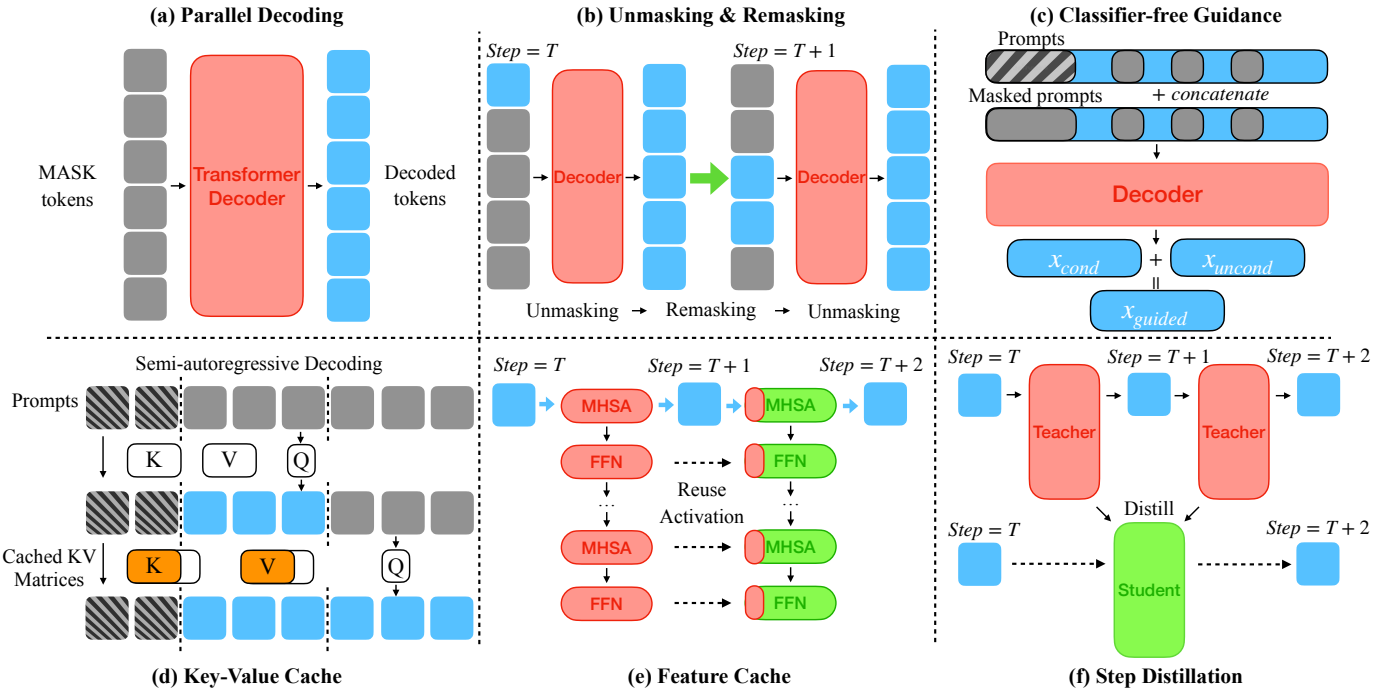

Fig. 5. 扩散语言模型的推理技术。我们在此展示六种不同的策略，包括：(a) 并行解码；(b) 去掩码和重掩码；(c) 无分类器引导；(d) 键值缓存；(e) 特征缓存；(f) 步骤蒸馏。

(2) 对立采样，在此情况下，相同的时间步长和掩码数据在当前策略 π_θ 和参考策略 π_{ref} 的 ELBO 估计中共享用于相同输入 y_w 或 y_l 。通过将 VRPO 应用于 LLaDA，所得的 LLaDA 1.5 模型在数学、代码和对齐基准测试中表现出显著且一致的改进。

4 推理策略

DLM 的推理策略有三个主要目标：(i) 通过解封和重新封装计划来提升生成质量，(ii) 实现更细致的内容控制，以及 (iii) 通过如 KV/特征缓存和步骤蒸馏等技术来提高效率。在图 5 中给出了一个简要概述。

4.1 并行解码

并行解码自然地与 DLMs 对齐，利用其固有的掩码预测能力来同时生成多个词元而不是顺序生成。然而，简单的并行化可能会降低连贯性，因此需要一系列适应性策略来平衡效率和质量。Fast-dLLM [40] 采用置信度感知解码，选择性地揭开预测概率超过某一阈值的词元的掩码，实现了高达 27.6 倍的速度提升而不降低质量。自适应并行解码（APD）[41] 通过咨询一个轻量级自回归辅助模型，动态调整并行度，在必要时在吞吐量和保真度之间进行权衡。SlowFast 采样 [82] 引入了一个两阶段的计划。首先是一个谨慎的“慢”阶段以定位稳定的词元，然后是一个积极的“快”阶段以批量最终确定它们，当与缓存结合使用时，实现高达 34 倍的加速。SpecDiff [74] 通过使用离散扩散模型作为一个完全并行的“起草者”，进一步推动吞吐量，其输出被一个更大的自回归模型快速验证（必要时校正），比传统的 AR 生成提高到 7.2 倍的加速。最后，视觉语言模型 Dimple [26] 采用置信度并行解码，动态调整每步揭示的词元数量，减少 1.5 - 7 倍的生成迭代。这些并行解码方法整体上显著缩小了扩散模型和自回归模型之间的延迟差距，同时在某些情况下保持甚至提高了生成质量。

先进的开源离散 DLMs（诸如 LLaDA [24] 和 Dream [22]）采用了一种 mask-predict 范式：在每个扩散步骤中，它们对高置信度的符号进行去掩码处理，并对不确定的部分重新加掩码，反复优化序列。因此，去掩码/重新掩码策略的选择，即低置信度采样、随机选择或自适应温度，主导了生成质量和收敛速度，使其成为最关键的推理杠杆之一。早期研究 Mask DLM [59] 形式化了两个基准：随机重掩码和置信度排序重掩码，显示出优先处理低置信度位置可以在不增加额外成本的情况下提高质量。基于这一洞察，Fast-dLLM [40] 引入了置信度感知的并行解码：每一步骤都对所有预测概率超过全局阈值的位置进行去掩码处理，实现了高达 13 倍的加速，同时保持准确度。最近，ReMDM [42] 提出了一种原则性推理时重新掩码采样器，能够对已解码的符号进行重新掩码以进一步优化；通过调整重掩码预算，它提供了顺滑的计算与质量折衷，并在固定计算条件下弥合了与自回归模型的质量差距。这些自适应去掩码/重新掩码策略共同显著提高了扩散语言模型的效率和质量，并可与稍后讨论的正交加速器（如缓存和步骤蒸馏）整合。

4.2 指导

引导是在扩散模型中一个关键的推理技术，通过将生成过程引导至期望的属性来提高输出的质量。在扩散模型中，引导是指任何改变模型去噪轨迹的技术，使得样本符合期望的条件，比如文本提示、类别标签或风格特征。这个想法由分类器引导 [13] 普及，即通过在得分估计中添加一个外部分类器的梯度，将样本推进到目标类别。不久之后，无分类器引导 [83] 去除了额外的分类器需求：模型通过有条件和无条件的方式训练一次，然后在推理时组合两个得分估计。其中 λ 是引导尺度，在对条件的保真度与样本多样性之间取得平衡。这一简单的公式现在支撑着大多数文本到图像系统（例如，Stable Diffusion [8] ）并已被 DLMs 用于提示控制生成。后续工作在几个方面完善了 CFG：使用 dropout 增强的 CFG 平滑了质量—多样性曲线；基于粒子的引导融合了多种条件；$p2$ 加权重新缩放了噪声项，以稳定高 λ 采样。在文本领域，更新的方案将引导扩展到结构和语义约束。FreeCache [84] 结合了

一种轻量级自回归验证器与一个离散 DLM：验证器在草稿标记被提交之前批准（或否决）它们，同时实施连贯性并启用激进的特征缓存。DINGO [85] 将正则表达式控制表述为 DFA 上的动态规划搜索，保证约束满足而不改变模型分布。在其他离散 DLMs 中，引导也可以在每个扩散步骤应用，可选地与掩码/重掩码或缓存结合，以引导内容（例如，主题、情感）同时保持效率。总体而言，引导已成为扩散推理的基石，提供了一种轻量、可调的手段，以便使模型输出与用户意图对齐。

4.3 高效推断

近期最先进的扩散语言模型 [22], [28], [81] 将经典的 Transformer 架构 [144] 与扩散过程的逐步随机推理过程相结合。因此，加速 DLM 推理的努力集中在两种互补的策略上：（1）降低 Transformer 框架每一步的计算开销，例如通过键值（KV）缓存或特征缓存。（2）减少扩散采样步骤的总数，例如通过步骤蒸馏。

传统的 KV 缓存利用 LLM 的严格自回归解码模式，因此不适合 DLM 的双向、多步生成范式。然而，最近的工作表明，精心重新设计解码计划可以恢复其大部分优势。Block Diffusion 引入了块离散去噪扩散语言模型（BD3-LMs），这些模型在每个块内运行扩散的同时，在粗略块之间以自回归方式解码文本；一旦一个块完成，其键和值被冻结并重用，能够实现可变长度的生成和可测量的速度提升。Fast-dLLM 保持块视图，但添加了一个无需训练的近似 DualCache，利用了前缀和后缀标记在连续扩散步骤中的 KV 激活的近似一致性，在 LLaDA 和 Dream 上达成高达 27× 的端到端吞吐量增益，同时仅有 $< 1\%$ 的准确性损失。补充了这些基于块的方案，dKV-Cache 观察到标记表示仅在位置解码后稳定，因此部署了一个延迟的条件缓存，将 KVs 存储延迟一步；这种设计在相同的模型上实现了 2 –10× 的加速且几乎没有质量下降。总体上，这些结果表明，半自回归调度和延迟缓存在扩散的双向条件和最初为自回归设计的 Transformer 技巧之间提供了实际的桥梁。

特征缓存首次由 DeepCache 引入，它利用中间 U-Net 激活在连续扩散步骤中强烈的相似性来避免冗余计算。后续工作 Δ -DiT、Learning-to-Cache 和 FasterCache 展示了同样的原理可以清晰地转移到基于 Transformer 的扩散模型上，实现了可比的加速效果而无需重新训练。随着扩散语言模型的兴起，dLLM-Cache 将特征缓存扩展到了文本，通过区分两种冗余：提示词在去噪过程中几乎保持静态，而响应词则仅稀疏变化。因此，它将一个长间隔的提示缓存与一个自适应短间隔的响应缓存配对，仅在轻量级价值相似性测试（“V-验证”）检测到明显变化时刷新，实现在 LLaDA-8B 和 Dream-7B 上高达 9× 的端到端加速。最近，FreeCache 缓存已经“清洁”的标记的 KV/特征投影，仅刷新动态位置，将加速进一步推至 34×，同时保持忠实度。总体而言，这些进展表明，特征缓存可以使扩散语言模型在推理延迟方面接近自回归 LLMs，而不牺牲输出质量。

Step Distillation. 步蒸馏是一种广泛采用的扩散模型加速技术，将通常的千步去噪过程压缩到仅几个甚至单个采样步骤，从而大幅减少推理时间。与之前讨论的不需要训练的方法不同，它施加了离线成本：必须首先训练一个紧凑的学生网络以模拟教师。早期的工作如渐进蒸馏 [91]，接着是 ADD [145] 和 LADD [146]，逐步减半步骤数或对齐中间分布以保持保真度。Di4C [90] 通过显式提取词间相关性，将框架扩展到离散扩散，使得学生在四到十步内即可匹配教师质量，同时实现 ~ 2 倍加速。最近，DLM-One [39] 使用基于分数的蒸馏与对抗正则化来训练一个连续扩散语言模型，可以在一个前向传递中生成整个序列，实现高达 500 倍的加速并接近教师质量。这些工作共同确立了步蒸馏作为缩小扩散和自回归语言模型之间时延差距的主要途径。

5 多模态和统一方法

本节探讨了将 DLM 扩展到多模态和统一架构的最新发展。与自回归 LLM 类似，DLM 可以自然地适应以处理多模态输入和输出。一种直接的方法是通过预训练的视觉编码器接受视觉输入。继 LLaVA [147] 在 AR 领域的成功之后，如 LLaDA-V [25]、LaViDa [92] 和 Dimple [26] 等模型使用视觉编码器来提取图像特征，然后将其投射到与文本标记相同的嵌入空间中。超越简单的视觉理解，DLM 为实现统一的多模态生成和理解提供了有希望的途径。由于共享的去噪扩散框架，DLM 自然支持不同模态的联合建模。视觉输入可以使用 VQ-VAE 离散化，从而在统一的标记空间中进行多模态输入和输出的训练。代表性模型如 MMaDA [27]、Fudoki [93] 和 Muddit [76] 都是这个方向的例子。我们首先介绍基于基础 LLaDA 模型的架构和预训练权重构建的 LLaDA LLaDA and LLaDA's Derivatives. 家族及其衍生模型。LLaDA-V [25] 集成了一个视觉编码器和一个基于 MLP 的投影器，该投影器将视觉特征映射到语言标记嵌入空间，从而实现有效的视觉指令调优。继承 LLaVA-NeXT [148]，LLaDA-V 采用了三阶段调优策略。在第一阶段，他们只训练 MLP 投影器，用 LLaVA 的训练数据将视觉表示与文本嵌入对齐。在第二阶段，模型通过使用 DLM 目标的大规模视觉指令数据 [149] 进一步调优。第三阶段通过在具有推理链的问答对上训练以增强多模态推理能力。虽然 LLaDA 的骨干在纯文本任务上略弱于 LLaMA3-8B [150]，但 LLaDA-V 在各种基准测试中实现了强大的性能和更好的可扩展性，与在相同数据上训练的 LLaMA3-V 相比。它缩小了与 Qwen2-VL [151] 的性能差距，并优于混合和纯 DLM 基础的模型 [75], [152], [153]，展示了扩散架构在多模态理解中的有效性。

LaViDa 引入了基于 LLaDA 和 Dream-7B 的 VLM 系列。通过使用预训练的视觉编码器，LaViDa 采用两阶段的训练策略分别训练投影器和微调模型。LaViDa 为解决多模态 DLM 的训练和推理挑战做出了显著贡献。通常，在掩码 DLM 中，平均只有约 50% 的标记被掩盖用于损失计算，这降低了效率，并可能在 VLM 训练中遗漏关键的答案标记，从而导致梯度失调。LaViDa 引入了互补掩码以实现有效训练：对于每个样本，生成两个具有不相交损坏范围的掩码版本，确保所有标记最终用于训练，提高了样本效率和梯度流。在推理过程中，LaViDa 利用 Prefix KV-Cache 缓存视觉和提示标记的键和值，大幅减少延迟，在性能仅有微小下降的情况下实现了最大 3.9 倍的加速。此外，时间步转移用于更早地去掩盖标记，进一步提升生成质量。实证结果表明，与基于 AR 的 VLM 相比，LaViDa 在推理速度显著加快的同时，取得了有竞争力或更优的性能。

在 LLaDA 的基础上，MMaDA [27] 进一步推广了架构，以支持多模态理解和生成。与之前的模型不同，MMaDA 通过使用 VQ-VAE 将图像标记化为离散代码，消除了对显式视觉编码器的需求，并通过模态无关的扩散变压器联合建模所有模态。该设计允许跨文本和图像模态的无缝集成，无需模态特定组件。MMaDA 还实施了一种混合长 CoT 微调策略，使 CoT 推理格式在模态之间对齐。此外，UniGRPO，一个基于政策梯度的统一 RL 算法，特别为扩散语言模型量身定制，使跨模态推理成为可能。不仅在文本推理方面超过了类似规模的模型如 LLaMA3 和在多模态理解方面超过了 Show-o [152]，MMaDA 甚至在图像生成上优于专业图像生成模型如 SDXL [10]。

Dimple . Dimple [26] 引入了一个大型多模态 DLM，将视觉编码器与离散 DLM 骨干结合。作者发现，纯粹的离散扩散训练方法存在明显的不稳定性、性能差以及严重的长度偏差问题。为了克服这些挑战，Dimple 提出了一种名为 Autoregressive-then-Diffusion 的新颖的两阶段训练范式。在

第一阶段，模型进行标准的自回归训练，以有效对齐视觉和语言模态。在第二阶段，它切换到基于扩散的训练，以恢复其并行解码能力。这种混合策略确保了稳定高效的训练，同时实现了与 LLaVA-NEXT 等当代自回归模型相当或更好的性能。

为了进行推理，Dimple 引入了几种技术来提高效率和可控性。自信解码根据信心阈值动态调整每一步生成的标记数，从而减少总的生成迭代次数。该模型还成功重新实现了自回归模型中常见的预填充技术，以缓存提示标记并实现高达 7 × 的加速，同时性能损失最小。此外，Dimple 探索了结构先验的使用，允许对响应格式和长度进行精确、细致的控制，这在自回归模型中是难以实现的功能。

D-DiT . 双重扩散变压器 (D-DiT) [75] 是一个大规模的全端到端统一多模态扩散模型，支持文本到图像（T2I）和图像到文本（I2T）任务。它直接解决了之前在视觉理解任务中面临的、主要由自回归模型主导的扩散模型的挑战。其结构灵感来源于多模态扩散变压器 (MM-DiT)，具有双分支变压器，可以处理图像和文本的标记，在每一层中使用注意力机制允许模态间的交互。该模型使用冻结的 VAE 进行图像处理和冻结的 T5 编码器进行文本处理，主要的骨干网络 MM-DiT 从预训练的 SD3 [11] 权重初始化。

D-DiT 的一个核心创新是其联合训练目标，该目标通过联合优化两种模态损失之和，将图像的连续潜空间扩散与文本的离散掩码标记扩散结合起来。与以前需要自回归组件来解码文本潜在变量的多模态扩散模型不同，D-DiT 完全基于扩散，并在与其他统一模型的竞争中表现出色。

UniDisc . 统一多模态离散扩散 (UniDisc) [94] 被提出作为一个统一的生成模型，用于联合文本和图像建模，基于离散扩散作为主流 AR 方法的替代方案。与先前讨论的 D-DiT 不同，UniDisc 在文本和图像标记上共同采用完整掩码的扩散过程，并具有全注意力机制，旨在学习将掩码标记的序列映射回来自共享词汇表的干净序列。训练是从头开始使用统一的离散扩散目标进行的，其中两个模态的标记被随机掩码，并且模型在监督下使用重新加权的交叉熵损失。

UniDisc 的一个关键优势是在条件生成任务中的卓越表现，这主要归功于对无分类器指导的有效利用。UniDisc 最显著的能力之一是能够以零样本的方式进行联合图像和文本修复，这是之前的 AR 或统一生成模型无法实现的。作者通过将模型扩大到 1.4B 进行规模分析，证明了 UniDisc 在性能和推理时间计算方面都优于 AR 模型，且具有增强的可控性和可编辑性。然而，发现 UniDisc 在达到相同的验证损失方面，其训练效率低于可比的 AR 模型。

Fudoki . Fudoki [93] 被介绍为第一个完全基于离散流匹配框架构建的通用统一多模态模型，挑战了自回归 (AR) 和基于掩码的扩散模型的主导地位。Fudoki 不依赖于简单掩码损坏过程，而是利用更加通用的度量诱导概率路径与动能最优速度，从而允许更具语义意义的损坏过程，并使模型能够在迭代细化过程中持续自我校正其预测。这种自我校正能力是与掩码的深度语言模型 (DLM) 的一个主要区别，其中未掩码的标记通常是固定的，无法进行修正。

为了减少从头开始训练的高成本，Fudoki 从一个预训练的基于自回归（AR）的多模态大模型（MLLM）Janus-1.5B [154] 初始化，然后通过一个两阶段过程适应离散流匹配范式。其架构基于 Janus-1.5B，但使用完整的注意力掩码以更好地捕捉全局语境，并移除时间嵌入层，因为模型可以从破损的输入中隐式推断时间步。Fudoki 在视觉理解和图像生成任务中达到了与最新的 AR 模型相当的性能，展示了推理速度和质量之间灵活的平衡。当应用测试时推理扩展技术时，模型性能显著提升，这表明该架构在下一代统一模型中具有进一步探索的潜力。

Muddit . Muddit [76] 是一个纯粹的统一离散扩散 Transformer，它将强大的文本到图像主干与轻量级的文本解码器集成在一起，在真正统一的架构下实现灵活和高质量的多模态生成。该模型从 Meissonoic [155] 预训练的 MM-DiT 初始化，使用统一的离散扩散目标进行训练，其中文本和图像标记根据余弦计划随机掩码，模型通过重新加权的交叉熵损失学习预测原始标记。通过语义丰富的视觉先验和并行离散扩散的结合，Muddit 在生成和理解基准上实现了与显著更大的 AR 模型相比竞争或更优的性能。它还展示了比 AR 基线快数倍的速度，突显了离散扩散方法在适当初始化时的效率和可扩展性。

在本节中，我们简要比较了各种 DLM 与 AR 模型的性能。我们展示了基于几个广泛使用的基准对 DLM 进行评估的可视化，包括用于普通语言理解的 PIQA [156] 和 HellaSwag [157]，用于代码生成的 HumanEval [158]，以及用于多模态生成和理解的 GenEval [159]、MME [160]、MMMU [161] 和 GQA [162]。我们还包括 GSM8K [163]，这是 DLM 文献中评估数学推理能力的一个流行基准。相应的性能可视化如图 5 所示。

调查的 DLMs 规模从不到 1B 到 8B 的参数不等。为了比较，我们还报告了相似规模的典型 AR 模型的性能。性能数据主要来自原始出版物。如果源论文中没有提供结果，我们参考了后续报道了类似评估的工作。

我们的研究结果表明，DLM 在与 AR 模型相当的规模下表现具有竞争力。在诸如 PIQA 和 HellaSwag 等一般语言理解基准上，像 LLaDA 这样的模型表现稍低于或与 AR 模型如 LLaMA2 [4] 和 Qwen2.5 [164] 持平。然而，在数学和科学相关的基准中，包括 GSM8K、GPQA [165] 和 MATH [166]，LLaDA 和 Dream 等模型始终优于同等大小的 AR 模型。在多模态任务中，像 MMaDA [27] 和 LLaDA-V [25] 这样的模型经常超越基于 AR 的多模态模型，突显了 DLM 在统一和跨模态推理中的潜力。在代码生成任务中，DLM 也展现了具有竞争力的能力。尤其是，DiffuCoder [80] 在开源模型中实现了具有竞争力的 HumanEval 表现，展示了 DLM 在结构化、逻辑复杂领域的潜力。同时，像 Gemini Diffusion [29] 和 Mercury [28] 这样的闭源 DLM 在所有 DLM 中达到最先进的结果，能够与顶级 AR 模型如 GPT-4o 媲美。

鉴于用于训练当前大多数 DLM 的训练数据和计算资源相对有限，这些结果表明 DLM 在许多实际应用中具有作为 AR 模型替代方案的强大潜力。

6 下游任务的应用

在大规模用于通用语言生成的 DLM 出现之前，DLM 已经被应用于各种传统 NLP 任务，如文本分类、命名实体/场景识别、情感分析、文档摘要、风格迁移、受限生成和机器翻译等。ROIC-DM 是第一个将扩散模型适应于鲁棒文本分类和推理的工作。它将扩散过程直接应用于类别标签，并在输入文本上对去噪过程进行条件化，这可以通过结合传统语言模型作为顾问进一步增强。DiffusionNER [96] 将命名实体识别定义为一个边界去噪任务。它将扩散过程应用于实体的开始和结束边界，通过迭代优化过程从随机噪声生成实体跨度。对于场景文本识别，IPAD [97] 引入了一种并行、迭代网络，将任务框架化为条件文本生成，采用离散扩散和易先解码方法有效平衡识别准确率和推理速度。对于基于方面的情感分析，DiffusionABSA [98] 使用扩散模型逐步提取方面。DiffuSum [99] 提出了一种用于抽取摘要的新范式，通过使用扩散模型直接生成所需的摘要句子表示。然后通过提取与这些生成的表示最佳匹配的文档句子形成最终摘要。对于法律文档摘要，TermDiffuSum [100] 提出了一种术语导向的扩散模型，通过多因素融合噪声加权计划优先考虑含有法律术语

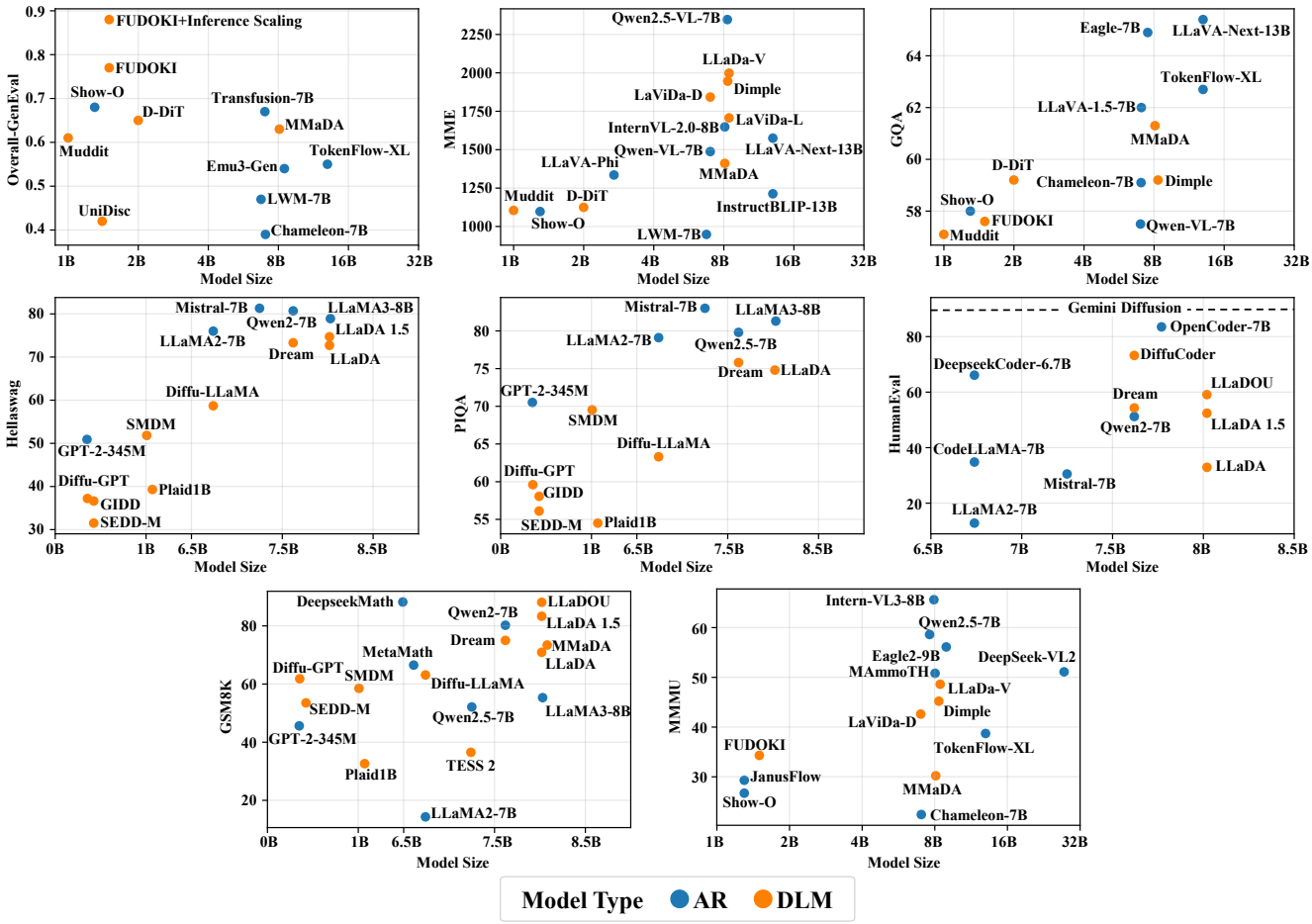

Fig. 6. **在八个基准测试上的性能比较**：Overall-GenEval、MME、CQA、Hellaswag、PIQA、HumanEval、GSM8K 和 MMMU。**每个子图中的横轴代表模型大小，以参数数量衡量。纵轴表示对应基准测试下的得分，得分越高表示性能越好。模型类型通过颜色区分**：**蓝色代表** AR **语言模型，而橙色代表** DLMs。

的句子。对于关键短语提取，Diff-KPE [101] 通过使用变分信息瓶颈引导文本扩散过程来生成并注入关键短语信息，增强短语表示。IPED [102] 将关系三元组提取视为隐式块扩散任务。EdiText [103] 通过结合基于 SDEdit 的技术与一种新颖的自我调节方法，引入了一种可控的由粗到细的文本编辑框架以实现精确的编辑控制。为了生成更具体的移情响应，DIFFUSEMP [104] 利用条件扩散模型，通过特殊的掩码策略综合多层次控制信号（例如，意图和语义框架）进行引导。DiffuDetox [105] 利用了一种混合扩散方法进行文本净化，结合了一个条件模型来降低毒性和一个无条件模型来确保输出文本的流畅性。一个微调的 DiffuSeq 模型被证明在细粒度文本风格转换任务中实现了最先进的性能 [167]，而 ParaGuide [106] 则引入了一个更灵活的即插即用框架，在推理时通过现成的分类器和风格嵌入器来指导释义条件扩散模型。为了生成流畅且多样化的段落，同时避免重复，PLANNER [107] 将潜在扩散规划模块用于生成语义段落嵌入，与自回归解码模块结合用于呈现最终文本。DiffuCom [108] 提出了一个高效的扩散模型用于评论生成，使用了上下文感知的注意力机制和自我条件化技术。DiffusionDialog [109] 通过在对话生成中执行扩散过程并使用连续潜变量来解决一对多问题，提高了响应的多样性和推理速度。对于释义生成，LDP [110] 在预训练模型的潜在空间中建模扩散，避免了通常的取整步骤以实现更高效能。对于高度受限的诗歌生成任务，PoetryDiffusion [111] 通过使用扩散模型生成语义而一个新颖的、独立训练的韵律控制器则强制执行结构规则如格式和韵律，独特地分离了任务。在机器翻译中，XDLM [112] 首创了一种跨语言预训练目标用于扩散模型，使其能够在预训练阶段有效学习语言之间的映射。DiffusionRet [113] 提出了一种两阶段生成检索方法，首先利用扩散模型从查询中生成伪文档，然后作为输入供基于 n 元模型检索最终文档。DIFND [114] 使用扩散模型生成反驳证据，并运用多代理 MLLM 系统进行链式反驳推理，以提高多模态假新闻检测的准确性和可解释性。

6.1 代码生成

虽然 DLMs 很少被明确设计用于代码生成，但它们的全局规划和迭代优化能力特别适合代码生成的非顺序特性。像 DiffuCoder [80] 这样的基础模型——一个 7B 开源模型，已经专门为这个领域开发。DiffuCoder 的分析揭示了一些独特的解码行为，例如在较高温度下生成顺序变得更加灵活。它还提出了 coupled-GRPO，一种新的采样方案，可以在训练中构建用于完成的互补掩膜噪声，从而显著提高模型在代码生成任务上的性能。在推理方面，DCoLT [78] 将整个反向扩散过程视为一种非线性“横向”思维。通过结果为导向的 RL 和去掩码策略模块，它在复杂的编码任务中取得了强劲的结果。Dilated Unmasking Scheduler (DUS) [115] 提供了一种仅推理的、无规划器的方法，在每次去噪步骤中以非相邻模式去掩码以最小化联合熵增益的上界，从而在代码生成上取得了有前景的结果，同时改善了速度质量的折衷。展示了 DLMs 的速度在现实世界中的潜力，Mercury Coder [28] 是一个商用规模的扩散模型，它在主要代码基准测试中保持相当的质

量的同时，在吞吐量方面达到了领先水平，性能比速度优化的自回归模型提升多达 10 倍。

6.2 生物学和科学应用

TransDLM [116] 通过目标特性文本描述引导的分子优化来避免误差传播。另一种文本引导的方法，TGM-DLM [117]，专注于通过集体和迭代更新 SMILES 字符串的标记嵌入来进行分子生成。无需依赖额外的数据资源，TGM-DLM 在生成性能上超越了 MolT5-Base。DRAKES [121] 为离散扩散模型引入了一种基于强化学习的微调方法，该方法使用 Gumbel-Softmax 技巧对 DNA 和蛋白质设计进行奖励反向传播。对于蛋白质建模，ForceGen [122] 通过使用蛋白质语言扩散模型生成满足复杂、非线性机械性能设计目标的序列，来实现全新的蛋白质设计。MeMDLM [118] 通过微调 ESM-2 蛋白质语言模型，引入了一种用于全新膜蛋白设计的掩码扩散语言模型，以生成新颖且逼真的跨膜序列。受 LLaDA 的启发，DSM [123] 引入了一种方案，同时实现高质量的表示学习和有效的生成性蛋白质设计。DPLM [119] 提供了一种多功能蛋白质语言模型，展示了强大的蛋白质序列生成和预测能力，并在表示学习中表现出色。DPLM2 [124] 进一步将模型扩展为多模态蛋白质基础模型，可以同时处理序列和结构。通过将三维结构坐标转换为离散标记，DPLM-2 学习这两种模态的联合分布。这不仅支持条件任务如蛋白质折叠和逆折叠，还能同时共生成兼容的蛋白质序列及其三维结构。CFP-GEN [120] 是一种新型的扩散语言模型，专为组合功能蛋白质生成而设计。它通过整合功能、序列和结构信息等多模态约束，促进全新的蛋白质设计。CFP-GEN 支持大规模高通量生成功能与天然蛋白质相当的新型蛋白质，并在多功能蛋白质设计中取得较高成功率。

7 挑战与未来方向

虽然扩散语言模型在广泛任务中表现出相当大的潜力，但仍有一些关键挑战存在，这限制了其实践部署和更广泛应用。在本节中，我们将概述并讨论需要进一步研究和创新的关键领域。

7.1 主要挑战

1) Parallelism–Performance Trade-off. 扩散语言模型旨在并行生成多个标记。然而，这种并行性通常以生成质量和一致性为代价。在离散的 DLM 中，同时在一个步骤中取消屏蔽多个标记会增加去噪的负担，这可能导致错误积累。一个核心问题是标记之间的相互依赖性，即所谓的并行解码诅咒 [40]。当同时预测多个标记时，模型为每个位置生成一个分布并独立地从中采样，未能考虑位置之间的依赖关系。考虑一个简单的例子，训练数据仅由两个序列组成：“ABABAB”和“BABABA”。统计上来说，“A”和“B”在每个位置的出现频率相同，这导致 DLM 在预测时赋予它们相似的概率。在自回归模型中，一旦生成第一个“A”，模型很可能会预测下一个为“B”，从而保持一致性。相比之下，一种并行生成标记的 DLM 可能会独立地为第一个和第二个位置采样“A”，生成类似“AAABBA”的序列，这偏离了有效的训练模式。实证研究表明，这一问题显著影响了 DLM 的性能，特别是在减少去噪步骤数量时 [23]。这一现象在图 7 中有所说明。未来的工作可能会集中于减轻这种权衡，潜在的方向包括引入结构化的约束，更明确地建模标记间的依赖性，或精炼采样策略以在并行生成期间改善连贯性。

2) Infrastructure. 虽然通过开源、高度优化的库和框架（例如，Hugging Face Transformers [168]）训练、微调和推理 AR 模型已经被显著简化和加速，但 DLM 在这方面仍然落后。目前，主要的机器学习生态系统几乎没有给予 DLM 本地支持，这给研究人员和开发者带来了实际的挑战。此外，在推理过程中，DLM 缺乏类似于 vLLM [169] 成熟的开源部署基础设施，使得高效服务 DLM 变得困难。

3) Long Sequence and Dynamic-Length Generation. DLMs 通常在基于扩散的目标下训练以对固定长度的序列进行去噪，这使得在推理时泛化到更长或动态大小的序列变得具有挑战性。大多数现有的 DLMs 限制在最大上下文长度为 4,096 个标记，并且在 DLM 环境中，对长序列使用的 AR 模型中的广泛使用的外推技术仍未被充分探索。这一限制阻碍了 DLMs 在需要长上下文理解或复杂推理的任务中的应用。此外，DLMs 通常要求在推理期间预先确定生成长度，这使它们不适合动态长度生成。尽管 DLMs 可以预测一个 [EOS] 标记并省略显示随后生成的标记，但在整个去噪过程中，整个序列仍然被完全更新，不管生成是否在逻辑上结束，这导致了不必要的计算开销。此外，掩码 DLMs 在每个去噪步骤中利用全双向注意力，这在每步中产生 $\mathcal{O}(N^2)$ 的计算成本，其中 N 是序列长度。假设每步有固定数量的标记被取消掩码，总去噪步骤的数量与 N 成线性关系，从而导致总体推理复杂度为 $\mathcal{O}(N^3)$。没有像 KV-Cache 这样的架构优化，这种立方时间复杂度严重限制了 DLMs 在真实世界应用中长序列生成的可扩展性。

扩展性仍然是扩散语言模型尚未充分研究的挑战，尤其是在与自回归模型相比时。虽然 DLMs 在某些指标和基准测试上显示出了有希望的结果，但它们的扩展程度仍未达到 AR 模型的水平。目前，公开可用的最大 DLM 仅包含约 80 亿个参数，显著小于已扩展到数百亿甚至上万亿的领先 AR 模型，例如 Llama-3.1-405B、DeepSeek-V3-671B-A37B MoE、Qwen3-235B-A22B MoE、Kimi-K2-1T-A32B MoE 等。闭源的 DLMs，如 Mercury 和 Gemini Diffusion，在广泛的基准测试中也未能达到最先进的 AR 模型的水平。此外，许多现有的 DLMs 要么是从先前预训练的 AR 模型中训练而来，要么是基于有限数据集构建在基础 DLMs（例如，LLaDA）之上，这进一步限制了它们的扩展性和性能。因此，进一步扩展 DLMs 的能力仍需验证或探索。

尽管存在上述挑战，但扩散语言模型（DLMs）在未来探索中展现出许多有前景的方向。以下，我们简要概述几个尚未充分探索的方向和机遇，这些方向和机遇可能显著推动该领域的发展：

在本次综述中，我们全面概述了扩散语言模型的整体布局。我们概述了 DLMs 的基本原理、分类和建模范式，并将它们与主流的自回归模型进行了比较，突出了它们的独特特点和优势。我们进一步探讨了训练和推理的设计空间，涵盖了质量和效率兼顾的各种训练策略和推理技术。此外，我们强调了多模扩散语言模型的最新进展，展示了其在处理多样数据模态方面的能力。最后，我们讨论了该领域的局限性和挑战，并指出了未来研究的有前景方向。我们希望这篇综述为对基于扩散的语言建模感兴趣的研究人员提供一个全面的参考，提供关于该领域现状及其未来前景的宝贵见解。我们也鼓励在这一令人振奋的研究领域进行进一步的探索和创新，因为扩散语言模型在不断发展，并推动语言理解和生成的边界。

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

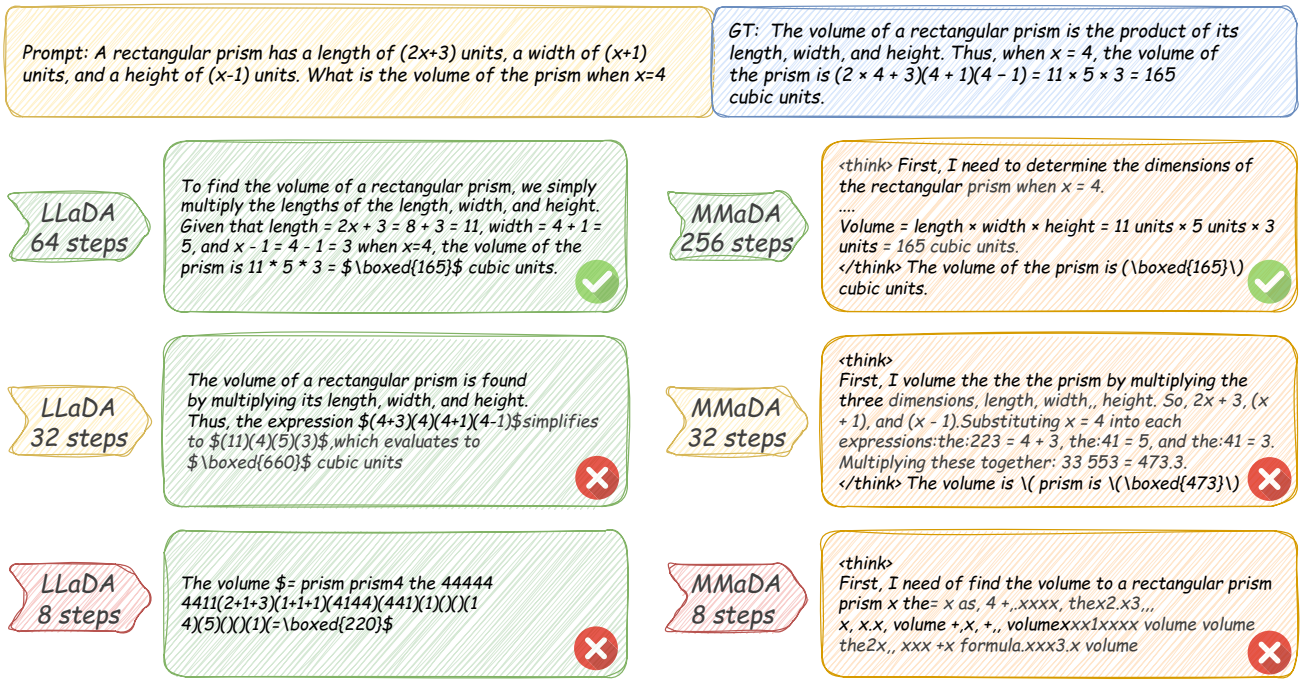

Fig. 7. 在不同去噪步骤设置下，LLaDA [24] 和 MMaDA [27] 的生成结果。注意，生成长度分别设置为 LLaDA 的 128 个标记和 MMaDA 的 256 个标记。仅当每步解码时有 1 或 2 个标记未被掩盖时，这两个模型才会生成正确且连贯的响应。在减少步骤和增加并行度的情况下，响应要么不正确，要么缺乏流畅性和一致性。这说明了在动态语言模型中并行性和输出质量之间的权衡。为了简化，我们省略了 MMaDA 在 256 步中的部分思考过程。

[3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[6] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.

[7] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[10] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*.

[11] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[12] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman *et al.*, "Video generation models as world simulators," *OpenAI Blog*, vol. 1, p. 8, 2024.

[13] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[15] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*.

[17] X. Liu, C. Gong *et al.*, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *The Eleventh International Conference on Learning Representations*.

[18] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Advances in neural information processing systems*, vol. 35, pp. 4328–4343, 2022.

[19] R. Strudel, C. Tallec, F. Altché, Y. Du, Y. Ganin, A. Mensch, W. Grathwohl, N. Savinov, S. Dieleman, L. Sifre *et al.*, "Self-conditioned embedding diffusion for text generation," *arXiv preprint arXiv:2211.04236*, 2022.

[20] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in neural information processing systems*, vol. 34, pp. 17 981–17 993, 2021.

[21] Z. He, T. Sun, Q. Tang, K. Wang, X. Huang, and X. Qiu, "Diffusionbert: Improving generative masked language models with diffusion models," in *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[22] J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong, "Dream 7b," 2025. [Online]. Available: https://hkunlp.github.io/blog/2025/dream

[23] S. Gong, S. Agarwal, Y. Zhang, J. Ye, L. Zheng, M. Li, C. An, P. Zhao, W. Bi, J. Han *et al.*, "Scaling diffusion language models via adaptation from autoregressive models," in *The Thirteenth International Conference on Learning Representations*.

[24] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," *arXiv preprint arXiv:2502.09992*, 2025.

[25] Z. You, S. Nie, X. Zhang, J. Hu, J. Zhou, Z. Lu, J.-R. Wen, and C. Li, "Llada-v: Large language diffusion models with visual instruction tuning," *arXiv preprint arXiv:2505.16933*, 2025.

[26] R. Yu, X. Ma, and X. Wang, "Dimple: Discrete diffusion multimodal large language model with parallel decoding," *arXiv preprint arXiv:2505.16990*, 2025.

[27] L. Yang, Y. Tian, B. Li, X. Zhang, K. Shen, Y. Tong, and M. Wang, "Mmada: Multimodal large diffusion language models," *arXiv preprint arXiv:2505.15809*, 2025.

[28] I. Labs, S. Khanna, S. Kharbanda, S. Li, H. Varma, E. Wang, S. Birnbaum, Z. Luo, Y. Miraoui, A. Palrecha *et al.*, "Mercury: Ultra-fast language models based on diffusion," *arXiv preprint arXiv:2506.17298*, 2025.

[29] DeepMind, "Gemini diffusion," https://deepmind.google/technologies/gemini, 2024, accessed: 2025-07-09.

[30] M. Xu, T. Geffner, K. Kreis, W. Nie, Y. Xu, J. Leskovec, S. Ermon, and A. Vahdat, "Energy-based diffusion language models for text generation," *arXiv preprint arXiv:2410.21357*, 2024.

[31] J. Deschenaux and C. Gulcehre, "Beyond autoregression: Fast llms via self-distillation through time," in *The Thirteenth International Conference on Learning Representations*.

[32] K. Han, K. Kenealy, A. Barua, N. Fiedel, and N. Constant, "Transfer learning for text diffusion models," *arXiv preprint arXiv:2401.17181*, 2024.

[33] S. S. Sahoo, J. Deschenaux, A. Gokaslan, G. Wang, J. Chiu, and V. Kuleshov, "The diffusion duality," *arXiv preprint arXiv:2506.10892*, 2025.

[34] Y. Zhang, S. He, D. Levine, L. Zhao, D. Zhang, S. A. Rizvi, E. Zappala, R. Ying, and D. van Dijk, "Non-markovian discrete diffusion with causal language models," *arXiv preprint arXiv:2502.09767*, 2025.

[35] M. Dang, J. Han, M. Xu, K. Xu, A. Srivastava, and S. Ermon, "Inference-time scaling of diffusion language models with particle gibbs sampling," *arXiv preprint arXiv:2507.08390*, 2025.

[36] L. Rout, C. Caramanis, and S. Shakkottai, "Anchored diffusion language model," *arXiv preprint arXiv:2505.18456*, 2025.

[37] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[38] S. Zhao, D. Gupta, Q. Zheng, and A. Grover, "d1: Scaling reasoning in diffusion large language models via reinforcement learning," *arXiv preprint arXiv:2504.12216*, 2025.

[39] T. Chen, S. Zhang, and M. Zhou, "Dlm-one: Diffusion language models for one-step sequence generation," *arXiv e-prints*, pp. arXiv–2506, 2025.

[40] C. Wu, H. Zhang, S. Xue, Z. Liu, S. Diao, L. Zhu, P. Luo, S. Han, and E. Xie, "Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding," *arXiv preprint arXiv:2505.22618*, 2025.

[41] D. Israel, G. V. d. Broeck, and A. Grover, "Accelerating diffusion llms via adaptive parallel decoding," *arXiv preprint arXiv:2506.00413*, 2025.

[42] G. Wang, Y. Schiff, S. S. Sahoo, and V. Kuleshov, "Remasking discrete diffusion models with inference-time scaling," in *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

[43] Z. Liu, Y. Yang, Y. Zhang, J. Chen, C. Zou, Q. Wei, S. Wang, and L. Zhang, "dllm-cache: Accelerating diffusion large language models with adaptive caching," *arXiv preprint arXiv:2506.06295*, 2025.

[44] X. Ma, R. Yu, G. Fang, and X. Wang, "dkv-cache: The cache for diffusion language models," *arXiv preprint arXiv:2505.15781*, 2025.

[45] G. Liu, Z. Feng, Y. Gao, Z. Yang, X. Liang, J. Bao, X. He, S. Cui, Z. Li, and Z. Hu, "Composable text controls in latent space with odes," *arXiv preprint arXiv:2208.00638*, 2022.

[46] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," in *The Eleventh International Conference on Learning Representations*.

[47] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan *et al.*, "Continuous diffusion for categorical data," *arXiv preprint arXiv:2211.15089*, 2022.

[48] Z. Gao, J. Guo, X. Tan, Y. Zhu, F. Zhang, J. Bian, and L. Xu, "Empowering diffusion models on the embedding space for text generation," *arXiv preprint arXiv:2212.09412*, 2022.

[49] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Q. Weinberger, "Latent diffusion for language generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 56 998–57 025, 2023.

[50] Z. Lin, Y. Gong, Y. Shen, T. Wu, Z. Fan, C. Lin, N. Duan, and W. Chen, "Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise," in *International Conference on Machine Learning*. PMLR, 2023, pp. 21 051–21 064.

[51] R. Wang, J. Li, and P. Li, "Infodiffusion: Information entropy aware diffusion process for non-autoregressive text generation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 13 757–13 770.

[52] G. Liu, Y. Wang, Z. Feng, Q. Wu, L. Tang, Y. Gao, Z. Li, S. Cui, J. Mcauley, Z. Yang *et al.*, "Unified generation, reconstruction, and representation: Generalized diffusion with adaptive latent encoding-decoding," in *International Conference on Machine Learning*. PMLR, 2024, pp. 31 964–31 993.

[53] A. Shabalin, V. Meshchaninov, and D. Vetrov, "Smoothie: Smoothing diffusion on token embeddings for text generation," *arXiv preprint arXiv:2505.18853*, 2025.

[54] R. K. Mahabadi, H. Ivison, J. Tae, J. Henderson, I. Beltagy, M. E. Peters, and A. Cohan, "Tess: Text-to-text self-conditioned simplex diffusion," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 2347–2361.

[55] J. Tae, H. Ivison, S. Kumar, and A. Cohan, "Tess 2: A large-scale generalist diffusion language model," *arXiv preprint arXiv:2502.13917*, 2025.

[56] P. Yu, S. Xie, X. Ma, B. Jia, B. Pang, R. Gao, Y. Zhu, S.-C. Zhu, and Y. N. Wu, "Latent diffusion energy-based model for interpretable text modelling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 702–25 720.

[57] L. Zheng, J. Yuan, L. Yu, and L. Kong, "A reparameterized discrete diffusion model for text generation," in *First Conference on Language Modeling*.

[58] J. Shi, K. Han, Z. Wang, A. Doucet, and M. Titsias, "Simplified and generalized masked diffusion for discrete data," *Advances in neural information processing systems*, vol. 37, pp. 103 131–103 167, 2024.

[59] S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. Chiu, A. Rush, and V. Kuleshov, "Simple and effective masked diffusion language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 130 136–130 184, 2024.

[60] J. Ye, Z. Zheng, Y. Bao, L. Qian, and Q. Gu, "Diffusion language models can perform many tasks with scaling and instruction-finetuning," *arXiv preprint arXiv:2308.12219*, 2023.

[61] K. Zhou, Y. Li, W. X. Zhao, and J.-R. Wen, "Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1438–1451.

[62] I. Gulrajani and T. B. Hashimoto, "Likelihood-based diffusion language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16 693–16 715, 2023.

[63] A. Lou, C. Meng, and S. Ermon, "Discrete diffusion modeling by estimating the ratios of the data distribution," in *International Conference on Machine Learning*. PMLR, 2024, pp. 32 819–32 848.

[64] J. Ou, S. Nie, K. Xue, F. Zhu, J. Sun, Z. Li, and C. Li, "Your absorbing discrete diffusion secretly models the conditional distributions of clean data," in *The Thirteenth International Conference on Learning Representations*, 2024.

[65] I. Gat, T. Remez, N. Shaul, F. Kreuk, R. T. Chen, G. Synnaeve, Y. Adi, and Y. Lipman, "Discrete flow matching," *Advances in Neural Information Processing Systems*, vol. 37, pp. 133 345–133 385, 2024.

[66] S. Liu, J. Nam, A. Campbell, H. Stark, Y. Xu, T. Jaakkola, and R. Gomez-Bombarelli, "Think while you generate: Discrete

diffusion with planned denoising," in *The Thirteenth International Conference on Learning Representations*, 2024.
[67] J. Ye, J. Gao, S. Gong, L. Zheng, X. Jiang, Z. Li, and L. Kong, "Beyond autoregression: Discrete diffusion for complex reasoning and planning," *arXiv preprint arXiv:2410.14157*, 2024.
[68] D. von Rütte, J. Fluri, Y. Ding, A. Orvieto, B. Schölkopf, and T. Hofmann, "Generalized interpolating discrete diffusion," in *Forty-second International Conference on Machine Learning*, 2025.
[69] X. Liu, Z. Liu, Z. Huang, Q. Guo, Z. He, and X. Qiu, "Longllada: Unlocking long context capabilities in diffusion llms," *arXiv preprint arXiv:2506.14429*, 2025.
[70] X. Han, S. Kumar, and Y. Tsvetkov, "Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 11 575–11 596.
[71] T. Wu, Z. Fan, X. Liu, H.-T. Zheng, Y. Gong, J. Jiao, J. Li, J. Guo, N. Duan, W. Chen *et al.*, "Ar-diffusion: Auto-regressive diffusion model for text generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39 957–39 974, 2023.
[72] M. Arriola, A. Gokaslan, J. T. Chiu, Z. Yang, Z. Qi, J. Han, S. S. Sahoo, and V. Kuleshov, "Block diffusion: Interpolating between autoregressive and diffusion language models," in *The Thirteenth International Conference on Learning Representations*.
[73] C. Huang and H. Tang, "Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation," *arXiv preprint arXiv:2505.14455*, 2025.
[74] J. K. Christopher, B. R. Bartoldson, T. Ben-Nun, M. Cardei, B. Kailkhura, and F. Fioretto, "Speculative diffusion decoding: Accelerating language generation through diffusion," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 12 042–12 059.
[75] Z. Li, H. Li, Y. Shi, A. B. Farimani, Y. Kluger, L. Yang, and P. Wang, "Dual diffusion for unified image generation and understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2779–2790.
[76] Q. Shi, J. Bai, Z. Zhao, W. Chai, K. Yu, J. Wu, S. Song, Y. Tong, X. Li, X. Li *et al.*, "Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model," *arXiv preprint arXiv:2505.23606*, 2025.
[77] J. Ye, S. Gong, L. Chen, L. Zheng, J. Gao, H. Shi, C. Wu, X. Jiang, Z. Li, W. Bi *et al.*, "Diffusion of thought: Chain-of-thought reasoning in diffusion language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105 345–105 374, 2024.
[78] Z. Huang, Z. Chen, Z. Wang, T. Li, and G.-J. Qi, "Reinforcing the diffusion chain of lateral thought with diffusion language models," *arXiv preprint arXiv:2505.10446*, 2025.
[79] O. Zekri and N. Boullé, "Fine-tuning discrete diffusion models with policy gradient methods," *arXiv preprint arXiv:2502.01384*, 2025.
[80] S. Gong, R. Zhang, H. Zheng, J. Gu, N. Jaitly, L. Kong, and Y. Zhang, "Diffucoder: Understanding and improving masked diffusion models for code generation," *arXiv preprint arXiv:2506.20639*, 2025.
[81] F. Zhu, R. Wang, S. Nie, X. Zhang, C. Wu, J. Hu, J. Zhou, J. Chen, Y. Lin, J.-R. Wen *et al.*, "Llada 1.5: Variance-reduced preference optimization for large language diffusion models," *arXiv preprint arXiv:2505.19223*, 2025.
[82] Q. Wei, Y. Zhang, Z. Liu, D. Liu, and L. Zhang, "Accelerating diffusion large language models with slowfast: The three golden principles," *arXiv preprint arXiv:2506.10848*, 2025.
[83] P. Li, S. Yan, J. Tsai, R. Zhang, R. An, Z. Guo, and X. Gao, "Adaptive classifier-free guidance via dynamic low-confidence masking," *arXiv preprint arXiv:2505.20199*, 2025.
[84] Z. Hu, J. Meng, Y. Akhauri, M. S. Abdelfattah, J.-s. Seo, Z. Zhang, and U. Gupta, "Accelerating diffusion language model inference via efficient kv caching and guided diffusion," *arXiv preprint arXiv:2505.21467*, 2025.
[85] T. Suresh, D. Banerjee, S. Ugare, S. Misailovic, and G. Singh, "Dingo: Constrained inference for diffusion llms," in *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.
[86] X. Ma, G. Fang, and X. Wang, "Deepcache: Accelerating diffusion models for free," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 762–15 772.
[87] P. Chen, M. Shen, P. Ye, J. Cao, C. Tu, C.-S. Bouganis, Y. Zhao, and T. Chen, "Δ-dit: A training-free acceleration method tailored for diffusion transformers," *arXiv preprint arXiv:2406.01125*, 2024.
[88] X. Ma, G. Fang, M. Bi Mi, and X. Wang, "Learning-to-cache: Accelerating diffusion transformer via layer caching," *Advances in Neural Information Processing Systems*, vol. 37, pp. 133 282–133 304, 2024.
[89] Z. Lv, C. Si, J. Song, Z. Yang, Y. Qiao, Z. Liu, and K.-Y. K. Wong, "Fastercache: Training-free video diffusion model acceleration with high quality," in *The Thirteenth International Conference on Learning Representations*.
[90] S. Hayakawa, Y. Takida, M. Imaizumi, H. Wakaki, and Y. Mitsufuji, "Distillation of discrete diffusion through dimensional correlations," in *Forty-second International Conference on Machine Learning*.
[91] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations*.
[92] S. Li, K. Kallidromitis, H. Bansal, A. Gokul, Y. Kato, K. Kozuka, J. Kuen, Z. Lin, K.-W. Chang, and A. Grover, "Lavida: A large diffusion language model for multimodal understanding," *arXiv preprint arXiv:2505.16839*, 2025.
[93] J. Wang, Y. Lai, A. Li, S. Zhang, J. Sun, N. Kang, C. Wu, Z. Li, and P. Luo, "Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities," *arXiv preprint arXiv:2505.20147*, 2025.
[94] A. Swerdlow, M. Prabhudesai, S. Gandhi, D. Pathak, and K. Fragkiadaki, "Unified multimodal discrete diffusion," *arXiv preprint arXiv:2503.20853*, 2025.
[95] S. Yuan, W. Yuan, H. Yin, and T. He, "Roic-dm: Robust text inference and classification via diffusion model," *arXiv preprint arXiv:2401.03514*, 2024.
[96] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Diffusionner: Boundary diffusion for named entity recognition," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 3875–3890.
[97] X. Yang, Z. Qiao, and Y. Zhou, "Ipad: Iterative, parallel, and diffusion-based network for scene text recognition," *International Journal of Computer Vision*, pp. 1–21, 2025.
[98] S. Liu, J. Zhou, Q. Zhu, Q. Chen, Q. Bai, J. Xiao, and L. He, "Let' s rectify step by step: Improving aspect-based sentiment analysis with diffusion models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 10 324–10 335.
[99] H. Zhang, X. Liu, and J. Zhang, "Diffusum: Generation enhanced extractive summarization with diffusion," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13 089–13 100.
[100] X. Dong, W. Li, Y. Le, Z. Jiang, J. Zhong, and Z. Wang, "Termdiffusum: a term-guided diffusion model for extractive summarization of legal documents," in *Proceedings of the 31st international conference on computational linguistics*, 2025, pp. 3222–3235.
[101] Y. Luo, Q. Zhou, and F. Zhou, "Enhancing phrase representation by information bottleneck guided text diffusion process for keyphrase extraction," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 6036–6047.
[102] J. Zhao, C. Xu, and B. Jiang, "Iped: An implicit perspective for relational triple extraction based on diffusion model," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 2080–2092.
[103] C. H. Lee, H. Kim, J. Yeom, and S. Yoon, "Editext: Controllable coarse-to-fine text editing with diffusion language models," *arXiv preprint arXiv:2502.19765*, 2025.
[104] G. Bi, L. Shen, Y. Cao, M. Chen, Y. Xie, Z. Lin, and X. He, "Diffusemp: A diffusion model-based framework with multi-grained

control for empathetic response generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 2812–2831.

[105] G. Floto, M. M. A. Pour, P. Farinneya, Z. Tang, A. Pesaranghader, M. Bharadwaj, and S. Sanner, "Diffudetox: A mixed diffusion model for text detoxification," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 7566–7574.

[106] Z. Horvitz, A. Patel, C. Callison-Burch, Z. Yu, and K. McKeown, "Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 16, 2024, pp. 18 216–18 224.

[107] Y. Zhang, J. Gu, Z. Wu, S. Zhai, J. Susskind, and N. Jaitly, "Planner: Generating diversified paragraph via latent language diffusion model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 178–80 190, 2023.

[108] J. Liu, P. Cheng, J. Dai, and J. Liu, "Diffucom: A novel diffusion model for comment generation," *Knowledge-Based Systems*, vol. 281, p. 111069, 2023.

[109] J. Xiang, Z. Liu, H. Liu, Y. Bai, J. Cheng, and W. Chen, "Diffusiondialog: A diffusion model for diverse dialog generation with latent space," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 4912–4921.

[110] W. Zou, Z. Zhuang, X. Geng, S. Huang, J. Liu, and J. Chen, "Improved paraphrase generation via controllable latent diffusion," *arXiv preprint arXiv:2404.08938*, 2024.

[111] Z. Hu, C. Liu, Y. Feng, A. T. Luu, and B. Hooi, "Poetrydiffusion: Towards joint semantic and metrical manipulation in poetry generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 279–18 288.

[112] L. Chen, A. Feng, B. Yang, and Z. Li, "Xdlm: Cross-lingual diffusion language model for machine translation," *arXiv preprint arXiv:2307.13560*, 2023.

[113] S. Qiao, X. Liu, and S.-H. Na, "Diffusionret: Diffusion-enhanced generative retriever using constrained decoding," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 9515–9529.

[114] K. Yan, M. Liu, Y. Liu, R. Fu, Z. Wen, J. Tao, and X. Liu, "Debunk and infer: Multimodal fake news detection via diffusion-generated evidence and llm reasoning," *arXiv preprint arXiv:2506.21557*, 2025.

[115] O. Luxembourg, H. Permuter, and E. Nachmani, "Plan for speed–dilated scheduling for masked diffusion language models," *arXiv preprint arXiv:2506.19037*, 2025.

[116] Y. Xiong, K. Li, J. Chen, H. Zhang, D. Lin, Y. Che, and W. Hu, "Text-guided multi-property molecular optimization with a diffusion language model," *arXiv preprint arXiv:2410.13597*, 2024.

[117] H. Gong, Q. Liu, S. Wu, and L. Wang, "Text-guided molecule generation with diffusion language model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 109–117.

[118] S. Goel, V. Thoutam, E. M. Marroquin, A. Gokaslan, A. Firouzbakht, S. Vincoff, V. Kuleshov, H. T. Kratochvil, and P. Chatterjee, "Memdlm: De novo membrane protein design with masked discrete diffusion protein language models," in *NeurIPS 2024 Workshop on AI for New Drug Modalities*.

[119] X. Wang, Z. Zheng, D. Xue, S. Huang, Q. Gu *et al.*, "Diffusion language models are versatile protein learners," in *Forty-first International Conference on Machine Learning*.

[120] J. Yin, C. Zha, W. He, C. Xu, and X. Gao, "Cfp-gen: Combinatorial functional protein generation via diffusion language models," in *Forty-second International Conference on Machine Learning*.

[121] C. Wang, M. Uehara, Y. He, A. Wang, A. Lal, T. Jaakkola, S. Levine, A. Regev, T. Biancalani *et al.*, "Fine-tuning discrete diffusion models via reward optimization with applications to dna and protein design," in *The Thirteenth International Conference on Learning Representations*.

[122] B. Ni, D. L. Kaplan, and M. J. Buehler, "Forcegen: End-to-end de novo protein generation based on nonlinear mechanical unfolding responses using a language diffusion model," *Science Advances*, vol. 10, no. 6, p. eadl4000, 2024.

[123] L. Hallee, N. Rafailidis, D. B. Bichara, and J. P. Gleghorn, "Diffusion sequence models for enhanced protein representation and generation," *arXiv preprint arXiv:2506.08293*, 2025.

[124] X. Wang, Z. Zheng, F. Ye, D. Xue, S. Huang, and Q. Gu, "Dplm-2: A multimodal diffusion protein language model," *arXiv preprint arXiv:2410.13782*, 2024.

[125] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[126] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[127] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*.

[128] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*.

[129] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[130] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[131] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[132] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[133] F. Gloeckle, B. Y. Idrissi, B. Roziere, D. Lopez-Paz, and G. Synnaeve, "Better & faster large language models via multi-token prediction," in *Forty-first International Conference on Machine Learning*.

[134] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[135] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[136] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[137] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[138] H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang, "Seqdiffuseq: Text diffusion with encoder-decoder transformers," *arXiv preprint arXiv:2212.10325*, 2022.

[139] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[140] Q. Team, "Qwen2.5: A party of foundation models," September 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

[141] X. Zhu, G. Karadzhov, C. Whitehouse, and A. Vlachos, "Segment-level diffusion: A framework for controllable long-form generation with diffusion language models," *arXiv preprint arXiv:2412.11333*, 2024.

[142] Y. Zihuiwen, Y. Elle Michelle, and B. Phil, "Latent diffusion for document generation with sequential decoding," in *NeurIPS 2023 Workshop on Diffusion Models*, 2023. [Online]. Available: https://neurips.cc/virtual/2023/74876

[143] M. Asada and M. Miwa, "Addressing the training-inference discrepancy in discrete diffusion for text generation," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 7156–7164.

[144] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[145] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[146] A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach, "Fast high-resolution image synthesis with latent adversarial diffusion distillation," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[147] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[148] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models," in *The Thirteenth International Conference on Learning Representations*.

[149] J. Guo, T. Zheng, Y. Bai, B. Li, Y. Wang, K. Zhu, Y. Li, G. Neubig, W. Chen, and X. Yue, "Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale," *arXiv preprint arXiv:2412.05237*, 2024.

[150] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[151] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[152] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou, "Show-o: One single transformer to unify multimodal understanding and generation," in *The Thirteenth International Conference on Learning Representations*.

[153] S. Kou, J. Jin, Z. Liu, C. Liu, Y. Ma, J. Jia, Q. Chen, P. Jiang, and Z. Deng, "Orthus: Autoregressive interleaved image-text generation with modality-specific heads," *arXiv preprint arXiv:2412.00127*, 2024.

[154] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 966–12 977.

[155] J. Bai, T. Ye, W. Chow, E. Song, Q.-G. Chen, X. Li, Z. Dong, L. Zhu, and S. Yan, "Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis," in *The Thirteenth International Conference on Learning Representations*, 2024.

[156] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.

[157] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4791–4800.

[158] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[159] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *Advances in Neural Information Processing Systems*, vol. 36, pp. 52 132–52 152, 2023.

[160] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[161] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.

[162] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[163] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[164] Q. Team, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[165] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q&a benchmark," in *First Conference on Language Modeling*, 2024.

[166] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[167] Y. Lyu, T. Luo, J. Shi, T. C. Hollon, and H. Lee, "Fine-grained text style transfer with diffusion-based language models," *arXiv preprint arXiv:2305.19512*, 2023.

[168] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[169] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.