arXiv:2603.01068v1 [cs.CV] 1 Mar 2026

LLaDA-o：一种高效且长度自适应的全模态扩散模型 el

Zebin You [1 2 3 †] **Xiaolu Zhang** [4] **JUN ZHOU** [4] **李崇轩** [1 2 3 ¶] **文继荣** [1 2 3]

摘要

我们提出了 **LLaDA-o**，一种用于多模态理解与生成的高效、长度自适应全模态扩散模型。LLaDA-o 构建于扩散混合框架（MoD）之上，将用于文本理解的离散掩码扩散与用于视觉生成的连续扩散解耦，同时通过一个共享、简洁且高效的注意力骨干网络将二者耦合，从而减少固定条件下的冗余计算。在 MoD 的基础上，我们进一步提出一种以数据为中心的长度适配策略，使模型无需改变架构，即可在多模态场景中灵活地进行不同长度的解码。大量实验表明，在多模态理解与生成基准测试中，LLaDA-o 的表现优于其他全模态扩散模型；在文本到图像生成的 DPG-Bench 上，其得分达到 87.04，证明了统一全模态扩散建模的有效性。代码见 https://github.com/ML-GSAI/LLaDA-o。

1. 引言

掩码扩散模型（MDMs）（Austin 等人，2021；Lou 等，2024；Shi 等，2024；Sahoo 等，2024；Ou 等，2024）近来已成为自回归（AR）语言模型的一种颇具前景的替代方案。通过并行迭代地对掩码词元进行去噪，掩码扩散模型展现出强大的语言建模性能（Nie 等，2025；Ye 等，2025；Cheng 等人，2025；Bie 等人，2025），并具备双向上下文建模和提升推理并行性等优势（Song 等人，2025；Khanna 等人，2025；谷歌，2025），因此越来越受到关注基于扩散的语言建模。

尽管取得了这一进展，扩散语言模型作为全模态模型在多模态理解与生成方面的潜力仍未得到充分探索。一个关键挑战在于，文本和图像偏好根本不同的扩散动态：掩码扩散天然适用于离散的语言词元，而对于图像而言，潜在空间中的连续扩散已成为事实标准（Sohl-Dickstein 等人，2015；Ho 等人，2020；Song 等人，2020b）。尽管已有相关尝试（Pynadath 等人，2025；Chang 等人，2022），但这些模态特定偏好无法直接迁移到不同领域，因此需要采用一种混合设计，对文本进行离散建模，对图像进行连续建模。

然而，在单个密集扩散模型中简单地联合训练两种模态，往往效果不佳。异构的状态空间和扰动过程可能导致目标不匹配与梯度干扰，进而造成优化不稳定和性能欠佳（Li 等人，2025c）。此外，现有的全模态扩散模型（Xin等人，2025；Li 等人，2025a；Yang 等人，2025a）通常假设理解任务的长度固定，从而限制了它们在开放式场景中的适用性。

为应对这些挑战，我们提出了**LLaDA-o**，这是一种基于*MixtureofDiffusion* (MoD)框架构建的高效长度自适应全模态扩散模型。MoD将适用于不同模态的扩散过程解耦为专用专家，同时保留共享的注意力骨干网络，以实现跨模态交互。具体而言，它为文本和视觉编码器词元分配掩码扩散，并为视觉潜在标记分配理解专家和生成专家，以缓解密集联合训练中的优化冲突。此外，我们在MoD的基础上进一步引入高效的注意力机制和以数据为中心的长度适配策略，从而在多模态场景中实现可扩展的推理和可变长度生成。

我们从定性和定量两个角度评估 LLaDA-o。定性方面，图1显示，与 Lumina-DiMOO 等现有全模态扩散模型相比，LLaDA-o 的多模态理解能力更强，生成的图像也包含更丰富的精细细节（Xin 等人，2025）。定量方面，我们在十项多模态理解基准上评估 LLaDA-o（见表1），结果表明，它在全模态扩散模型中取得了最先进的成果，并

†在蚂蚁集团实习期间完成的工作 [1]中国人民大学高瓴人工智能学院，北京，中国。[2]北京大模型与智能治理重点实验室。[3]下一代智能搜索与推荐工程研究中心，教育部。[4]蚂蚁集团。通讯作者：李崇轩 <chongxuanli@ruc.edu.cn>。

预印本。*2026*年*3*月*3*日。



图 *1.* **LLaDA-o 功能概览**。 上：多模态理解示例。中：遵循复杂提示词的文本到图像生成结果（提示词见表 7）。下：与现有全能扩散模型的逐案例比较。LLaDA-o 展现出更强的理解能力，并能遵循指令生成精细细节更丰富的图像。

其他基于离散扩散的多模态方法。我们还在广泛使用的 GenEval（Ghosh等人，2023）和 DPG-Bench（Hu等人，2024）基准测试上，将文本到图像生成效果与强大的纯生成模型和统一多模态模型进行比较。LLaDA-o 在 DPG-Bench 上取得了最先进的表现（87.04）。此外，我们还对 LLaDA-o 进行了全面分析。总体而言，这些结果证明了 LLaDA-o 作为统一全模态扩散模型的有效性。

2. 预备知识

我们介绍扩散模型的预备知识。

2.1. 连续扩散模型

连续扩散模型（CDM）（Sohl-Dickstein 等，2015；Ho 等人，2020；Song 等人，2020b）是现代图像生成的核心范式，尤其是在与扩散 Transformer（Peebles & Xie，2023；Bao 等人，2023）结合时。从概念上讲，CDM 定义了一个前向随机过程，逐步将数据扰乱为噪声，并学习逆转这一过程，以从噪声分布中生成样本。这一机制可以用随机微分方程来刻画（Song 等人，2020b）。在推理时，同一动力学过程可表示为等价的常微分方程（ODE）公式，从而得到确定性采样器（Song 等人，2020a；Lu 等人，2022）。这一 ODE 视角还启发了一系列变体，包括整流流（RF）（Liu，2022）和流匹配（Lipman 等人，2022）。

特别地，RF（Liu，2022）将加噪后的分布π_0 与数据分布 π_1 通过确定性的线性路径连接起来。给定一对数据与噪声（$\boldsymbol{x}_0$, $\boldsymbol{x}_1$），时间 $t \in [0,1]$ 处的中间状态定义为严格的线性插值：

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1. \tag{1}$$

RF 通过最小化以下目标函数，学习速度场 $p_{\boldsymbol{\theta}}(\boldsymbol{x},\ t)$，使其与恒定流动方向 $\boldsymbol{x}_1 - \boldsymbol{x}_0$ 一致：

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{x}_1),t}\Big[\big\|(\boldsymbol{x}_1-\boldsymbol{x}_0) - p_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\big\|_2^2\Big]. \tag{2}$$

采样对应于求解 ODE $d\boldsymbol{z}_t/dt\ = p_{\boldsymbol{\theta}}(\boldsymbol{z}_t,\ t)$，其初始值为 $\boldsymbol{z}_0 \sim \pi_0$。形式上，RF 对时间区间上的学习到的速度场进行积分，并将其离散化（例如，使用欧拉方法）：

$$\boldsymbol{z}_1 = \boldsymbol{z}_0 + \int_0^1 p_{\boldsymbol{\theta}}(\boldsymbol{z}_t,t)\,dt. \tag{3}$$

2.2. 扩散语言模型

离散扩散模型（Sohl-Dickstein 等人，2015；Austin 等人，2021）是专为离散数据（如词元序列）设计的生成模型。与 CDMs 类似，它们定义了一个离散状态的前向马尔可夫过程，逐步将数据破坏至简单先验，通常是词元上的均匀分布或吸收掩码状态。随后，模型学习相应的反向动态，以便从该先验生成离散数据。离散扩散的一种特例——掩码扩散模型（MDMs）（Lou 等人，2024；Shi 等人，2024；Sahoo 等人，2024；Ou 等人，2024），已在语言建模中展现出强大潜力与良好的缩放特性。

形式上，令 $\boldsymbol{x}_0 = [\boldsymbol{x}^i]_{i=1}^N$ 表示由 N 词元组成的序列，[M] 表示一个特殊掩码标记。在前向过程中，根据从 $[0,1]$ 均匀采样的时间步 t，独立地对各词元进行扰动。令 α_t 为连续单调递增函数，且 $\alpha_0 = 0$ 和 $\alpha_1 = 1$。每个词元以 α_t 的概率保持不变，或以 $1-\alpha_t$ 的概率替换为 [M]，如下所示：

$$\begin{aligned} q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) &= \prod_{i=0}^{N-1} q_{t|0}(\boldsymbol{x}_t^i|\boldsymbol{x}_0^i), \\ q_{t|0}(\boldsymbol{x}_t^i|\boldsymbol{x}_0^i) &= \begin{cases} \alpha_t, & \boldsymbol{x}_t^i = \boldsymbol{x}_0^i, \\ 1-\alpha_t, & \boldsymbol{x}_t^i = \text{[M]}. \end{cases} \end{aligned} \tag{4}$$

反向过程会以迭代方式重建序列。具体而言，考虑从时间 t 到s 的变换（其中 $0 \le s < t \le 1$）。已经被预测的词元保持不变。相反，掩码词元要么以 $\frac{1-\alpha_s}{1-\alpha_t}$ 的概率继续保持掩码状态，要么根据模型预测 $p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t)$ 以 $\frac{\alpha_s-\alpha_t}{1-\alpha_t}$ 的概率进行解码：

$$\begin{aligned} q_{s|t}(\boldsymbol{x}_s|\boldsymbol{x}_t) &= \prod_{i=0}^{N-1} q_{s|t}(\boldsymbol{x}_s^i|\boldsymbol{x}_t), \\ q_{s|t}(\boldsymbol{x}_s^i|\boldsymbol{x}_t) &= \begin{cases} 1, & \boldsymbol{x}_t^i \neq \text{[M]}, \boldsymbol{x}_s^i = \boldsymbol{x}_t^i, \\ \frac{1-\alpha_s}{1-\alpha_t}, & \boldsymbol{x}_t^i = \text{[M]}, \boldsymbol{x}_s^i = \text{[M]}, \\ \frac{\alpha_s-\alpha_t}{1-\alpha_t} p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t), & \boldsymbol{x}_t^i = \text{[M]}, \boldsymbol{x}_s^i \neq \text{[M]}, \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{5}$$

在推理过程中，掩码扩散模型通过迭代模拟这一反向转移来生成文本，逐步将完全掩码的序列转换为连贯的文本。

直观而言，掩码扩散模型充当掩码预测器，旨在根据观测上下文恢复掩码词元。形式上，掩码扩散模型的训练目标定义为：

$$\mathcal{L}(\boldsymbol{\theta}) = \int_0^1 \frac{1}{t}\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t\sim q_{t|0}}\Big[\sum_{i:\boldsymbol{x}_t^i=\text{[M]}} -\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t)\Big]dt. \tag{6}$$

扩散大语言模型（dLLMs）（Nie 等人，2025；Ye 等人，2025）证明了掩码扩散技术在大规模应用中的可行性，在实现具有竞争力的性能的同时，还支持并行解码和灵活的生成控制。

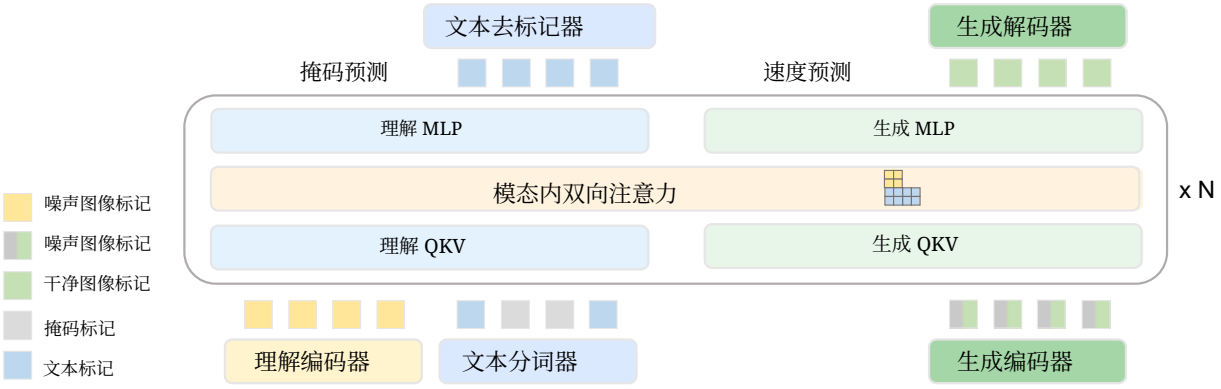


图 2. **LLaDA-o 概览：扩散混合框架。**

3. 方法

受第 1 节讨论的启发，**LLaDA-o** 在单一扩散框架内统一了多模态理解与生成：将离散模态与连续模态分离，再通过共享且高效的注意力机制将二者耦合，并支持自适应长度训练与推理。具体如下。

3.1. 扩散混合框架

LLaDA-o 的核心是一种混合扩散设计，它对离散词元和连续视觉潜在表示采用各自最优的参数化形式。然而，在单个稠密模型中直接联合训练这些模态往往效果不佳（Li 等人，2025c）：这两个分支运行于异构状态空间，并采用不同的扰动过程，可能导致目标不匹配和梯度干扰，进而引发训练冲突并使性能不理想。

为应对这些挑战，我们提出了扩散混合（MoD），这是一种统一的多模态扩散框架，如图2所示。受模态因子化设计的启发（Liang 等人，2024；Deng等人，2025），MoD 使用两个扩散专家，将离散模态和连续模态的处理解耦：一个理解专家通过掩码扩散处理文本和视觉编码器词元；另一个生成专家通过连续扩散处理视觉潜在标记。虽然这两个专家各司其职，但它们共用同一个自注意力骨干网络（见第3.2节），以确保有效的跨模态交互。

具体而言，理解专家依次集成了一个视觉编码器（Zhai 等人，2023）、一个轻量级两层多层感知机，以及一个扩散语言模型（Nie 等人，2025）。编码器将图像编码为语义视觉标记，多层感知机将其投影到语言标记空间，再由语言模型与提示标记共同处理。给定一个训练样本$(\boldsymbol{v},\ \boldsymbol{p},\ \boldsymbol{r}_0)$，其中$\boldsymbol{v}$ 表示投影后的图像词元，$\boldsymbol{p}$ 表示提示词，而 $\boldsymbol{r}_0$ 表示真实响应，整个专家经过训练，以优化式（6）的变体，如下所示：

$$\mathcal{L}_{\text{und}} = \int_0^1 \frac{1}{t}\mathbb{E}_{\substack{\boldsymbol{v},\boldsymbol{p}\\ \boldsymbol{r}_0,\boldsymbol{r}_t}}\left[\sum_{i:\boldsymbol{r}_t^i=[\text{M}]} -\log p_{\boldsymbol{\theta}}(\boldsymbol{r}_0^i|\boldsymbol{v},\boldsymbol{p},\boldsymbol{r}_t)\right]dt. \tag{7}$$

与此同时，生成专家由一个变分自编码器 (VAE) (Kingma& Welling，2013) 和一个扩散 Transformer (Peebles &Xie，2023) 组成。VAE 在图像与视觉潜在标记之间进行映射，其参数在训练期间保持冻结。给定相关变量 $(\boldsymbol{p},\ \boldsymbol{v}_0,\ \boldsymbol{\epsilon})$，其中 $\boldsymbol{p}$ 表示提示词，$\boldsymbol{v}_0$ 表示来自 VAE 的真实图像词元，而 $\boldsymbol{\epsilon}$ 表示高斯噪声，训练目标遵循公式 (2)：

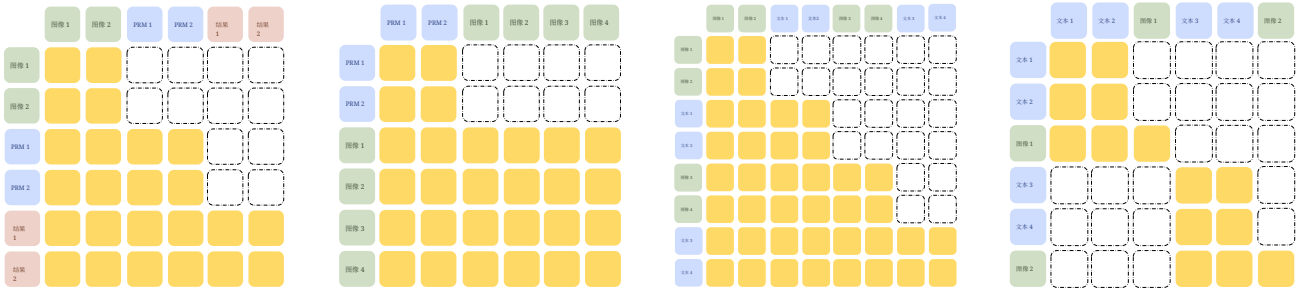
$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\boldsymbol{p},\boldsymbol{\epsilon},\boldsymbol{v}_0,\boldsymbol{v}_t,t}\left[\left\|(\boldsymbol{v}_0-\boldsymbol{\epsilon})-p_{\boldsymbol{\theta}}(\boldsymbol{p},\boldsymbol{v}_t,t)\right\|_2^2\right]. \tag{8}$$

值得注意的是，生成任务中的输入图像和文本也都由理解专家处理。因此，相应的参数会联合训练。多轮对话和交错数据的详细训练方案见附录 B.2。

3.2. 模态内双向注意力

尽管 MoD 通过不同的专家实现了模态特定处理的解耦，但这些专家共享一个用于跨模态交互的注意力骨干网络。在这种设置下使用全局注意力效率低下，因为它会在每个去噪步骤中重新计算整个序列上的注意力，即使条件（例如输入图像或文本提示词）保持不变也是如此。

为此，我们提出模态内双向注意力，一种简单而高效、专为 MoD 定制的注意力方案。具体而言，我们将输入序列划分为模态块，在每个块内应用全注意力，并在块之间施加因果注意力。该设计在保留丰富模态内上下文的同时实现高效推理：条件块构成固定前缀，其



(a) 多模态理解 *(b)*多模态生成 *(c)* 交错数据 *(d)* 跨样本隔离

图 *3.* **模态内双向注意力的实现**。 黄色块表示未掩码 注意力，而虚线白框表示已掩码 注意力。在情形 (a–b) 中，文本序列被明确划分为提示（PRM）和响应（RES）。

键值（KV）缓存可在各去噪步骤之间复用，从而避免在长序列上的冗余计算。

所提出的注意力方案可通过在模态块之间引入适当的注意力掩码来实现。如图3（a–c）所示，该原则自然适用于不同数据类型，包括多模态理解、生成和交错数据。此外，我们采用样本打包，并在样本之间进行严格隔离（见图 3d），以防止跨样本干扰。

与具有全局双向注意力的代表性基线（You 等人，2025a）相比，我们的注意力设计在实践中实现了 5.9 倍加速（见图 4）。

3.3. 自适应长度增强

理解专家采用掩码扩散进行文本建模，通常假设推理时目标长度固定。为实现灵活生成，同时避免引入架构修改或破坏样本隔离，我们设计了一种以数据为中心的策略，称为 自适应长度增强，它与我们的注意力方案完全兼容。

如算法 1所示，在训练期间，每个独立样本中的目标响应会以两种互补的方式进行随机扰动。以 p_{ext} 的概率，在原始响应后追加有限数量的 [序列结束符] 词元，使模型能够接触到不同位置的显式终止。以 p_{trunc} 的概率，将响应截断为随机前缀，促使模型学会从部分目标中正确续写。重要的是，这两种操作都保持严格的样本隔离，无需修改注意力结构或序列打包策略。

在推理时（参见附录 B.5中的算法2），我们在模态内双向注意力下进行分块生成（Arriola 等人，2025）：固定的条件块（图像和提示词）只缓存一次，并通过追加长度为L的掩码块并对其去噪，迭代生成响应词元。如果 [序列结束符] 以较高置信度出现，则解码终止；否则，已完成的块会被缓存，生成过程继续进入下一个块。这种策略能够高效地进行可变长度生成，同时充分复用固定前缀的 KV 缓存。

与以往依赖架构修改或采用全局双向注意力的多样本打包来处理可变长度输出的方法相比（Wu 等人，2025d；Kim 等人，2025；Yang 等人，2025b），我们的方法依然轻量、与架构无关，并且天然兼容样本隔离和高效推理。

算法 1 自适应长度训练

输入： 模型参数 $\boldsymbol{\theta}$；训练数据 $(\boldsymbol{v},\ \boldsymbol{p},\ \boldsymbol{r}_0)$；扩展概率 p_{ext}；截断概率 p_{trunc}。 **重复**: $\tilde{\boldsymbol{r}}_0 \leftarrow \boldsymbol{r}_0$, $u \sim \mathcal{U}(0,\ 1)$ **如果** $u < p_{\text{ext}}$ **则**: 在 $(1,\ |\boldsymbol{r}_0|)$ 之间均匀采样一个整数 k，并通过追加 k 个 EOS 令牌来更新 $\tilde{\boldsymbol{r}}_0$。 **否则如果** $u < p_{\text{ext}} + p_{\text{trunc}}$ **且** $|\boldsymbol{r}_0| > 16$ **则**: 在 $(1,\ |\boldsymbol{r}_0|-1)$ 之间均匀采样一个整数 ℓ，并将其截断至前 ℓ 个词元以更新 $\tilde{\boldsymbol{r}}_0$。 **结束如果** 使用 $(\boldsymbol{v},\ \boldsymbol{p},\ \tilde{\boldsymbol{r}}_0)$ 进行训练，仅在响应令牌上计算损失 $\tilde{\boldsymbol{r}}_0$；更新 $\boldsymbol{\theta}$ **直到**收敛

4. 实验

我们介绍实验设置、结果及分析。

4.1. 实验设置

模型。对于理解专家，我们采用具有代表性的 LLaDA-8B-Instruct（Nie 等人，2025）来初始化语言模型。视觉编码器采用 SigLIP（Zhai 等人，2023），它在许多多模态大语言模型中表现出色；投影器则采用随机初始化的两层多层感知机。对于生成专家，我们采用与掩码预测器架构相同的扩散 Transformer 架构，

表 *1.* **多模态理解基准测试上的评估**。 符号 † 表示 LaViDa-O（Li 等人，2025a）的结果，而 ⋆ 表示我们使用官方检查点和推理脚本评估的结果。“-”表示缺失数据，“Diff.”指扩散语言模型。值得注意的是，在基于扩散的统一多模态模型中，LLaDA-o 实现了最先进的性能。

模型	MMMU	MME		SeedB	MMB	MathVerse	MathVista	AI2D	ChartQA	DocVQA	InfoVQA
	val	cog.	困惑度	图像	en-dev	mini-vis。	testmini	-	-	val	val
基于自回归											
Emu3（Wang 等人，2024）	31.6	-	-	68.2	58.5	-	-	70.0	68.6	76.3	43.8
Janus-Pro（Chen等人，2025c）	41.0	-	1567	72.1	79.2	-	-	-	-	-	-
MetaMorph（Tong 等人，2025）	41.8	-	-	71.8	75.2	-	-	-	37.1	-	-
Show-o（Xie等人，2024）	27.4	-	1232	-	-	-	-	-	-	-	-
Show-o2（Xie等人，2025）	48.9	-	1620	-	79.3	-	-	78.6	-	-	-
BAGEL（Deng等人，2025）	**55.3**	-	**1687**	-	**85.0**	-	**73.1**	-	-	-	-
基于扩散											
LaViDa-L（Li 等人，2025b）	43.3	341	1365	-	70.5	27.2	44.8	70.0	64.6	59.0	34.2
Dimple(Yu 等人, 2025)	45.2	432	1514	-	74.6	-	42.3	74.4	63.4	-	-
LLaDA-V (You 等人, 2025a)	48.6	491	1507	74.8	82.9	28.5	59.7	77.8	78.3	83.9	**66.3**
MMaDA(Yang 等人, 2025a)	30.2	242†	1410	64.2	68.5	13.5†	33.7†	66.6†	9.8†	10.9†	14.9†
Lumina-DiMOO(Xin等人，2025）	**58.6**	-	**1534**	**83.1**	**84.5**	10.3⋆	30.3⋆	43.2⋆	8.3⋆	7.2⋆	6.2⋆
LaViDa-O (Li 等人，2025a)	45.1	488	1431	-	76.4	36.9	56.9	76.7	80.0	73.7	44.6
LLaDA-o	44.9	**549**	1412	75.3	71.1	**37.1**	**66.1**	**79.3**	**87.9**	**91.5**	54.7

该架构见于 LLaDA（Nie 等人，2025），并以其为基础进行初始化；时间嵌入的额外条件参数则采用随机初始化。由于 FLUX（Labs，2024）的变分自编码器具有出色的重建质量，因此我们将其用作生成任务的视觉编码器。

训练策略。我们分三个阶段训练 LLaDA-o，逐步提升数据难度和生成保真度。阶段 1 中，我们使用大规模图像理解数据和图像生成数据，并将生成限制在不超过 512 的分辨率，以稳定训练。阶段 2 中，我们引入多模态推理数据，并重复使用阶段 1 图像生成数据中的高质量子集，同时将生成分辨率提高到 1024，以提升高分辨率合成效果。值得注意的是，在前两个阶段，我们不会对多模态理解应用自适应长度增强。阶段 3 中，我们联合应用自适应长度增强，为理解专家启用可变长度生成，并加入更多高质量图像生成数据，使模型能够灵活地进行文本解码并实现更强的视觉生成。有关训练策略和数据的更多详情，请参阅附录 B.1。

评估。 我们在一系列广泛的基准上评估 LLaDA-o，以反映统一多模态模型的主要要求：通识理解、推理和细粒度感知，以及图像生成。对于多模态理解，我们涵盖了多学科知识（MMMU (Yue等人，2024)、MME (Fu等人，2023)、SEED-Bench (Li 等人，2023) 和 MM-Bench (Liu等人，2024)）、数学推理（Math-Verse (Zhang等人，2024) 和 MathVista (Lu等人，2023)），以及图表/文档理解（AI2D (Kembhavi等人，2016)、ChartQA (Masry等人，2022)、DocVQA (Mathew 等人，2021) 和 InfoVQA (Mathew等人，2022)）。对于文本到图像生成，我们采用两个广泛使用的基准，它们用于测试生成的互补方面：GenEval (Ghosh等人，2023)，该基准通过以物体为中心的检测流程（例如，物体数量、空间关系和颜色绑定）验证细粒度组合属性；以及 DPG-Bench (Hu等人，2024)，该基准评估对包含复杂实体关系和丰富描述的长篇、高信息密度提示词的忠实呈现。

4.2. 基准测试结果

多模态理解。 我们在表格中将 LLaDA-o 与统一多模态模型和多模态大语言模型进行比较。1。值得注意的是，LLaDA-o 在全模态扩散模型中取得了最先进的 表现（例如 LaViDa-O（Li 等人，2025a）），展现了扩散混合的有效性。这一优势在数学推理（例如 MathVista（Lu 等人，2023））以及图表/文档理解（例如 ChartQA（Masry 等人，2022））方面尤为明显。这些提升证明了我们的 MoD 框架行之有效。与最先进的自回归模型 BAGEL（Deng等人，2025）相比，LLaDA-o 整体表现较弱，这在预期之中，因为 BAGEL 的语言骨干网络更强。BAGEL 使用 Qwen2.5-7B-Instruct（Team，2024b）进行训练，使用了 18T 个词元；而我们的 LLaDA-8B-Instruct（Nie 等人，2025）使用了 2.3T 个词元进行训练，这一差距也体现在语言能力上（例如，HumanEval 得分为 84.8 对 49.4）。尽管处于劣势，LLaDA-o 仍缩小了差距；例如，在 MathVista 上，其得分达到 66.1，优于 LLaDA-V（You 等，2025a）（59.7），并接近 BAGEL（73.1）。我们认为，随着掩码扩散骨干网络持续改进，MoD 也将进一步提升。

表2. **在GenEval基准上对文本到图像生成的评估**（Ghosh等人, 2023）。“Gen.”、“Obj.”和“Attr.”分别表示生成、物体和属性，而“-”表示缺失数据。与最先进的统一多模态模型相比，LLaDA-o展现出更优的性能，尤其是在双物体和特定颜色生成任务上。遵循BAGEL（Deng等人, 2025）和Show-o2（Xie等人, 2025）的协议，我们使用改写提示词评估结果。

模型	# 参数	单物体	双物体	计数	颜色	位置	颜色属性	总体↑
仅生成								
PixArt-α (Chen等人，2023)	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
DALL-E 3 (Betker 等人，2023)	-	0.96	0.87	0.47	0.83	**0.43**	0.45	0.67
SD3-Medium(Esser 等人，2024)	2B	**0.99**	**0.94**	0.72	**0.89**	0.33	**0.60**	**0.74**
FLUX.1-dev (Labs, 2024)	12B	0.98	0.81	**0.74**	0.79	0.22	0.45	0.66
统一								
Emu3 (Wang 等人，2024)	8B	-	-	-	-	-	-	0.66
Janus（Wu 等人，2025a）	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus-Pro（Chen等人，2025c）	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Mogao（Liao 等人，2025）	7B	**1.00**	0.97	0.83	0.93	0.84	0.80	0.89
MMaDA（Yang 等人，2025a）	8B	0.99	0.76	0.61	0.84	0.20	0.37	0.63
Show-o（Xie等人，2024）	1.3B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
BAGEL（Deng等人，2025）	70亿 MoT	0.98	0.95	0.84	0.95	0.78	0.77	0.88
LaViDa-O (Li 等人，2025a)	-	-	-	-	-	-	-	0.89
Lumina-DiMOO (Xin 等人，2025)	8B	**1.00**	0.96	**0.87**	0.95	**0.85**	0.82	**0.91**
Show-o2 (Xie等人，2025)	7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76
LLaDA-o（本文方法）	8B MoT	0.99	**0.98**	0.73	**0.96**	0.69	**0.83**	0.86

表3. **在 DPG-Bench 上对文本到图像生成进行**（Hu等人，2024）的评估。符号 † 表示来自 Lumina- DiMOO 的结果（Xin等人，2025），“Gen.”表示生成，“-”表示缺少数据。值得注意的是，与先前仅支持生成的模型和统一模型相比，LLaDA-o 达到了最先进的性能。

模型	参数量	全局	实体	属性	关系	其他	总体↑
仅生成							
PixArt-α (Chen等人，2023)	0.6B	74.97	79.32	78.60	82.57	76.96	71.11
PixArt-Σ (Chen等人，2024)	0.6B	86.89	82.89	88.94	86.59	87.68	80.54
DALL-E 3（Betker 等人，2023）	-	**90.97**	89.61	88.39	90.58	**89.83**	83.50
SD3-Medium（Esser 等人，2024）	2B	87.90	**91.01**	88.83	80.70	88.68	**84.08**
FLUX.1-dev（Labs，2024）	12B	74.35	90.00	**88.96**	**90.87**	88.33	83.84
统一							
Emu3-DPO（Wang 等人，2024）	8B	-	-	-	-	-	81.60
Janus（Wu 等人，2025a）	1.3B	82.33	87.38	87.70	85.46	86.41	79.68
Janus-Pro（Chen等人，2025c）	7B	86.90	88.90	89.40	89.32	89.48	84.19
Mogao（Liao 等人，2025）	7B	82.37	90.03	88.26	93.18	85.40	84.33
MMaDA† （Yang 等人，2025a）	8B	77.81	78.48	81.74	84.79	63.20	69.97
Show-o（Xie等人，2024）	1.3B	-	-	-	-	-	67.48
BAGEL（Deng等人，2025）	70亿 MoT	88.94	90.37	**91.29**	90.82	88.67	85.07
LaViDa-O(Li 等人，2025a)	-	-	-	-	-	-	83.20
Lumina-DiMOO (Xin 等人，2025)	8B	81.46	92.08	88.98	**94.31**	82.00	86.04
Show-o2(Xie等人，2025)	7B	89.00	91.78	89.96	91.81	91.64	86.14
LLaDA-o（我们的方法）	8B MoT	**92.91**	**93.30**	90.40	91.75	**92.79**	**87.04**

文生图生成。 我们在 GenEval（Ghosh等人，2023）和 DPG-Bench（Hu等人，2024）（表格 2和 3）上，将 LLaDA-o 与最先进的纯生成模型及统一多模态模型进行评估。在 GenEval 上，LLaDA-o 的表现优于 Janus 等大多数强劲模型——Pro（Chen等人，2025c）和 SD3-Medium（Esser 等人，2024）。尽管整体表现略逊于 Lumina-DiMOO（Xin 等人，2025）和 Mogao（Liao 等人，2025），但它在双物体生成和颜色绑定方面表现更佳。值得注意的是，在 DPG-Bench 上，LLaDA-o 达到了最先进的

表 4. **置信度阈值对 MathVista 的影响**。我们报告了不同置信度阈值下的准确率（%）和吞吐量（词元/秒），块长度固定为 64。

置信度阈值（r）	0.2	0.4	0.6	0.8	0.9	1.0
准确率 (%)	57.8	62.4	64.2	65.0	65.9	65.8
吞吐量（词元/秒）	203.9	134.5	97.3	64.1	52.2	24.3

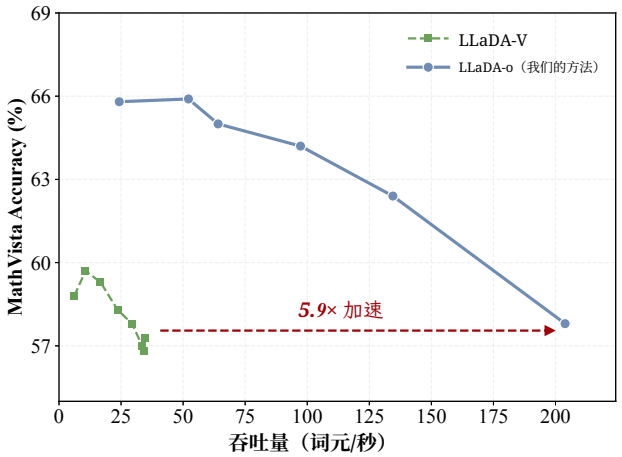


图 4. **MathVista 上推理效率的比较**。我们通过调整 LLaDA-o 的置信度阈值以及应用于 LLaDA-V 的 Fast-dLLM 的刷新间隔（n），来展示吞吐量与准确率之间的权衡。我们的方法优于 LLaDA-V，在性能相当的情况下实现了 5.9× 加速。

性能（87.04），超过了 Show-o2（Xie等人，2025）和 Lumina-DiMOO，表明其对于信息密集型长提示词具有出色的生成质量。这些结果证明了我们的 MoD框架将连续扩散模型与基于dLLM的骨干网络相结合以实现统一多模态生成的有效性。从定性结果来看，图 1 显示，与 Lumina-DiMOO 和 LaViDa-O（Li 等人，2025a）相比，LLaDA-o 在遵循指令的同时生成了视觉效果更佳、精细细节更丰富的图像。

我们还在附录中提供了更多定性文生图样本。C.1。

4.3. 进一步分析

表4 和图4 分析了推理效率和生成质量。与自回归模型不同，LLaDA-o 可以通过置信度阈值灵活调节这一权衡：提高阈值，通过仅选择高置信度词元来优先保证准确率；降低阈值则能加快生成速度。实验证明，阈值设为 0.9 时可实现最佳平衡。与 MathVista 上最先进的基线方法进行比较时（图4），LLaDA-o 展现出显著的效率提升。最值得注意的是，LLaDA-o 实现了 5.9× 相较于 LLaDA-V 的加速。这一显著改进验证了我们提出的模态内双向注意力能够减少计算冗余并实现高效推理。

表5。 **块长度对 MathVista 的影响**。我们报告不同块长度下生成词元的平均数量，置信度阈值固定为 0.95。

块长度	32	64	96	128
平均词元数	165	148	145	154
准确率（%）	63.6	66.1	66.2	65.3

表6。 **不同训练阶段的文本到图像生成性能**。我们报告 GenEval 和 DPG-Bench 上的评估结果，以展示性能的提升过程。在此表中，我们使用 GenEval 的原始提示词。

训练阶段	GenEval	DPG-Bench
阶段 1	0.74	86.1
阶段 2	0.78	86.2
阶段 3	**0.82**	**87.0**

我们在附录 C.2、C.3以及表5中对变长行为进行了全面分析。$L \in \{16, \ldots, 128\}$定性来看，与 LLaDA-V 在块设置不匹配时会生成冗余或不完整文本不同，LLaDA-o 能够生成长度恰当的内容，且在不同块大小下保持一致（$L \in \{16, \ldots, 128\}$）。定量来看，表 5显示，生成长度相对稳定：将块长度从 32 增加到 96，平均词元数量仅略微下降（从 165 降至 145），而准确率则有所提升（从 63.6% 升至 66.2%）。这些结果表明，输出长度主要由输入内容决定，而非预设的块大小，证实了我们所提出的自适应长度增强的有效性。

最后，表 6 研究了各训练阶段对文本到图像生成的影响。GenEval 和 DPG-Bench 上的结果从阶段 1 到阶段 3 持续提升，其中阶段 3 表现最佳（0.82 和 87.0），这证明了我们的多阶段训练流程是有效的。为完整起见，我们在附录 B.4中报告了三阶段训练流程的计算成本。

5. 结论

我们提出了 LLaDA-o，这是一种用于多模态理解与生成的长度自适应全模态扩散模型。LLaDA-o 基于扩散混合框架，采用共享的高效注意力骨干网络和以数据为中心的自适应长度训练策略，实现了稳定的多模态训练和可变长度生成。实验结果表明，该模型在多模态理解和文本到图像生成任务上表现出色。我们相信，随着掩码扩散模型在语言建模领域不断发展，LLaDA-o 有望为未来的全模态扩散方法提供坚实的基础。

参考文献

An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., 等。Llava-onevision-1.5：面向普及多模态训练的全开放框架。*arXiv* 预印本 *arXiv:2509.23661*，2025。

Arriola, M., Gokaslan, A., Chiu, J. T., Yang, Z., Qi, Z., Han, J., Sahoo, S. S., and Kuleshov, V. 块扩散：在自回归语言模型与扩散语言模型之间进行插值。*arXiv* 预印本 *arXiv:2503.09573*，2025。

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. 离散状态空间中的结构化去噪扩散模型。神经信息处理系统进展，34:17981–17993，2021。

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. 所有词语皆有价值：一种用于扩散模型的 ViT 主干网络。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第22669–22679页，2023。

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y. 等。使用更好的图像说明改进图像生成。计算机科学。*https://cdn. openai. com/papers/dall-e-3. pdf*，2(3)：8，2023。

Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z. 等。Llada2. 0：将扩散语言模型扩展至100b。*arXiv*预印本 *arXiv:2512.15745*，2025。

Chang, H., Zhang, H., Jiang, L., Liu, C. 和 Freeman, W. T. MaskGit：掩码生成式图像变换器。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第11315–11325页，2022。

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., 等。Pixart-α：面向逼真文本到图像合成的扩散Transformer快速训练。*arXiv* 预印本 *arXiv:2310.00426*，2023。

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., 和 Li, Z。Pixart-σ：面向 4k 文本到图像生成的扩散Transformer弱到强训练。载于欧洲计算机视觉会议，第 74–91 页。施普林格，2024。

Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., 和 Wang, B。Sharegpt-4o-image：使多模态模型与 GPT-4o 级图像生成对齐。*arXiv* 预印本 *arXiv:2506.18095*，2025a。

Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S. 等。BLIP3-o：一系列完全开放的统一多模态模型——架构、训练与数据集。*arXiv* 预印本 *arXiv:2505.09568*，2025b。Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X. 和 Ruan, C. Janus-Pro：基于数据与模型缩放的统一多模态理解与生成。*arXiv* 预印本 *arXiv:2501.17811*，2025c。

Chen, Z., Bai, X., Shi, Y., Fu, C., Zhang, H., Wang, H., Sun, X., Zhang, Z., Wang, L., Zhang, Y., 等。Opengpt-4o-image：用于高级图像生成与编辑的综合数据集。*arXiv* 预印本 *arXiv:2509.24900*, 2025d。

Cheng, S., Bian, Y., Liu, D., Zhang, L., Yao, Q., Tian, Z., Wang, W., Guo, Q., Chen, K., Qi, B., 等。SDAR：一种用于可扩展序列生成的协同扩散自回归范式。*arXiv* 预印本 *arXiv:2510.06303*, 2025。

Deng, C., Zhu, D., Li, K., Gou,C., Li, F.,Wang, Z.,Zhong, S., Yu, W., Nie, X., Song, Z., 等。统一多模态预训练中的新兴特性。*arXiv* 预印本 *arXiv:2505.14683*, 2025。

Esser, P., Kulal, S., Blattmann, A., Entezari, R., M¨uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., 等。用于高分辨率图像合成的整流流Transformer扩展。发表于第四十一届国际机器学习会议，2024。

Fang, R., Yu, A., Duan, C., Huang, L., Bai, S., Cai, Y., Wang, K., Liu, S., Liu, X., 和 Li, H. FLUX-Reason-6M 与 prism-bench：百万规模的文生图推理数据集及综合基准。*arXiv*预印本 *arXiv:2509.09680*，2025。

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., 和 Ji, R. MME：面向多模态大语言模型的综合评估基准。*arXiv*预印本 *arXiv:2306.13394*，2023。

Ghosh, D., Hajishirzi, H. 和 Schmidt, L. GenEval：一种用于评估文生图对齐的以物体为中心的框架。神经信息处理系统进展，36:52132–52152，2023。

谷歌。GeminiDiffusion，2025。网址 https://deepmind.google/models/gemini-diffusion。

Gu, S., Zhang, J., Zhou, S., Yu, K., Xing, Z., Wang, L., Cao, Z., Jia, J., Zhang, Z., Wang, Y., 等。Infinity-mm：借助大规模、高质量指令数据提升多模态性能。*arXiv*预印本 *arXiv:2410.18558*，2024。

Guo, J., Zheng, T., Li, Y., Bai, Y., Li, B., Wang, Y., Zhu, K., Neubig, G., Chen, W., and Yue, X. Mammoth-vl：通过大规模指令微调激发多模态推理。载于计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 13869–13920 页，2025。

Ho, J.、Jain, A. 和 Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020。

Hoogeboom, E.、Nielsen, D.、Jaini, P.、Forr´e, P. 和 Welling, M. Argmax流与多项扩散：学习分类分布。*NeurIPS*，34:12454–12465，2021。

Hu, X.、Wang, R.、Fang, Y.、Fu, B.、Cheng, P. 和 Yu, G. Ella：为扩散模型配备大语言模型，以增强语义对齐。*arXiv* 预印本 *arXiv:2403.05135*，2024。

jackyhate. jackyhate/text-to-image-2m。 https://huggingface.co/datasets/jackyhate/text-to-image-2M, 2024。

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H. 和 Farhadi, A. 一张图胜过十几张图像。载于 计算机视觉–*ECCV 2016*：第 *14* 届欧洲计算机视觉会议，阿姆斯特丹，荷兰，*2016* 年 *10* 月 *11–14* 日，论文集，第 *IV* 部分 *14*，第 235–251 页。施普林格，2016 年。

Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., Ermon, S. 等。Mercury：基于扩散的超高速语言模型。*arXiv* 预印本 *arXiv:2506.17298*，1，2025 年。

Kim, J., Cheuk-Kit, L., Domingo-Enrich, C., Du, Y., Kakade, S., Ngotiaoco, T., Chen, S. 和 Albergo, M. 任意顺序灵活长度掩码扩散。*arXiv* 预印本 *arXiv:2509.01025*，2025 年。

Kingma, D. P. 和 Welling, M. 自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013。

Labs，B.F. Flux。 https://github.com/black-forest-labs/flux，2024。

Li, B.、Wang, R.、Wang, G.、Ge, Y.、Ge, Y. 和 Shan, Y. SEED-Bench：使用生成式理解对多模态大语言模型进行基准测试。*arXiv* 预印本 *arXiv:2307.16125*，2023。

Li, S., Gu, J., Liu, K., Lin, Z., Wei, Z., Grover, A. 和 Kuen, J. LaViDa-O：面向统一多模态理解与生成的弹性大型掩码扩散模型。*arXiv* 预印本 *arXiv:2509.19244*，2025a。

Li, S., Kallidromitis, K., Bansal, H., Gokul, A., Kato, Y., Kozuka, K., Kuen, J., Lin, Z., Chang, K.-W. 和 Grover, A. LaViDa：面向多模态理解的大型扩散语言模型。*arXiv* 预印本 *arXiv:2505.16839*，2025b。Li, Z., Li, H., Shi, Y., Farimani, A. B., Kluger, Y., Yang, L. 和 Wang, P. 面向统一图像生成与理解的双重扩散。载于计算机视觉与模式识别会议论文集，第 2779–2790 页，2025c。Liang, W., Yu, L., Luo, L., Iyer, S., Dong, N., Zhou, C., Ghosh, G., Lewis, M., Yih, W.-t., Zettlemoyer, L. 等。变换器混合：一种面向多模态基础模型的稀疏且可扩展架构。*arXiv* 预印本 *arXiv:2411.04996*，2024。

Liang, Z.、Li, Y.、Yang, T.、Wu, C.、Mao, S.、Nian, T.、Pei, L.、Zhou, S.、Yang, X.、Pang, J. 等。离散扩散视觉-语言-动作：将离散扩散引入视觉-语言-动作策略中的动作解码。*arXiv* 预印本 *arXiv:2508.20072*，2025。

Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., 和 Huang, W. Mogao：一种用于交错式多模态生成的全能基础模型。*arXiv* 预印本 *arXiv:2505.05472*，2025。

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., 和 Le, M. 用于生成建模的流匹配。*arXiv* 预印本 *arXiv:2210.02747*，2022。

Liu, Q. 整流流：一种保持边缘分布的最优传输方法。*arXiv* 预印本 *arXiv:2209.14577*，2022。

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., 等。Mmbench：你的多模态模型是全能选手吗？载于欧洲计算机视觉会议，第 216–233 页。施普林格，2024。

Lou, A.、Meng, C. 与 Ermon, S. 通过估计数据分布的比率进行离散扩散建模。载于第四十一届国际机器学习会议，2024。

Lu, C.、Zhou, Y.、Bao, F.、Chen, J.、Li, C. 与 Zhu, J. Dpm-solver：一种约 10 步完成扩散概率模型采样的快速常微分方程求解器。神经信息处理系统进展，35:5775–5787，2022。

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Math-vista：使用 GPT-4V、Bard 和其他大型多模态模型评估视觉情境中的数学推理。*CoRR*，2023。

Ma, J., Zhu, X., Pan, Z., Peng, Q., Guo, X., Chen, C., and Lu, H. X2Edit：通过自构建数据和任务感知表征学习重新审视任意指令图像编辑。*arXiv* 预印本 *arXiv:2508.07607*，2025。Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. ChartQA：一个关于图表问答的基准，涵盖视觉推理和逻辑推理。*arXiv* 预印本 *arXiv:2203.10244*，2022。Mathew, M., Karatzas, D., and Jawahar, C. DocVQA：一个用于文档图像视觉问答的数据集。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第 2200–2209 页，2021。

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. InfographicVQA。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第 1697–1706 页，2022。

Meyer, J., Padgett, N., Miller, C., and Exline, L. 公共领域 12m：一个具有新颖治理机制的高美学图文数据集。*arXiv* 预印本 *arXiv:2410.23144*，2024。

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. 大型语言扩散模型。*arXiv* 预印本 *arXiv:2502.09992*，2025。

Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., 和 Li, C. 你的吸收离散扩散实际上对干净数据的条件分布进行了建模。*arXiv* 预印本 *arXiv:2406.03736*，2024。

Ou, J., Han, J., Xu, M., Xu, S., Xie, J., Ermon, S., Wu, Y., 和 Li, C. 基于序列级视角的扩散大语言模型原则性强化学习。*arXiv* 预印本 *arXiv:2512.03759*，2025。

Peebles, W. 和 Xie, S. 基于 Transformer 的可扩展扩散模型。发表于*IEEE/CVF* 计算机视觉国际会议论文集，第 4195–4205 页，2023。

Pynadath, P., Shi, J., 和 Zhang, R. Candi：混合离散-连续扩散模型。*arXiv* 预印本 *arXiv:2510.22510*，2025。

Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A., and Kuleshov, V. 简单且有效的掩码扩散语言模型。*arXiv* 预印本 *arXiv:2406.07524*，2024。

Shi, J., Han, K., Wang, Z., Doucet, A., 和 Titsias, M. K. 面向离散数据的简化且通用的掩码扩散。*arXiv*预印本 *arXiv:2406.04329*，2024。

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., 和 Ganguli, S. 使用非平衡热力学进行深度无监督学习。发表于国际机器学习会议，第 2256–2265 页。PMLR，2015。Song, J., Meng, C., 和 Ermon, S. 去噪扩散隐式模型。*arXiv* 预印本 *arXiv:2010.02502*，2020a。Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Er- mon, S., 和 Poole, B. 通过随机微分方程进行基于分数的生成建模。*arXiv* 预印本 *arXiv:2011.13456*，2020b。

Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., 等。Seed 扩散：一种具备高速推理能力的大规模扩散语言模型。*arXiv* 预印本 *arXiv:2508.02193*，2025。

Swerdlow, A., Prabhudesai, M., Gandhi, S., Pathak, D., and Fragkiadaki, K. 统一多模态离散扩散。*arXiv* 预印本 *arXiv:2503.20853*，2025。

Team, C. Chameleon：混合模态早期融合基础模型。*arXiv* 预印本 *arXiv:2405.09818*，2024a。

Team, Q. Qwen2.5：基础模型家族，2024 年 9 月 2024b。URL https://qwenlm.github.io/blog/qwen2.5/。

Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., 和 Liu, Z. MetaMorph：通过指令微调实现多模态理解与生成。发表于*IEEE/CVF*国际计算机视觉会议论文集，第 17001–17012 页，2025。

Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q. 等。Emu3：你所需的一切就是下一词元预测。*arXiv* 预印本 *arXiv:2409.18869*，2024。

Wen, J., Zhu, M., Liu, J., Liu, Z., Yang, Y., Zhang, L., Zhang, S., Zhu, Y., and Xu, Y. dvla：结合多模态思维链的扩散视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2509.25681*，2025a。

Wen, Y., Li, H., Gu, K., Zhao, Y., Wang, T., and Sun, X. Llada-vla：视觉语言扩散动作模型。*arXiv* 预印本 *arXiv:2509.06932*，2025b。

Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al. Janus：面向统一多模态理解与生成的视觉编码解耦。载于计算机视觉与模式识别会议论文集，第 12966–12977 页，2025a。

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., 等。Qwen-Image 技术报告。*arXiv* 预印本 *arXiv:2508.02324*，2025b。

Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., 等。OmniGen2：迈向先进多模态生成的探索。*arXiv* 预印本*arXiv:2506.18871*，2025c。

Wu, Z., Zheng, L., Xie, Z., Ye, J., Gao, J., Feng, Y., Li, Z., W., V., Zhou, G. 和 Kong, L. Dreamon：面向超越固定尺寸画布的代码填充的扩散语言模型，2025d。
U
RLhttps://hkunlp.github.io/blog/2025/dreamon。

Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z. 和 Shou, M. Z. Show-o：用单个 Transformer 统一多模态理解与生成。*arXiv* 预印本 *arXiv:2408.12528*，2024。

Xie, J., Yang, Z. 和 Shou, M. Z. Show-o2：改进的原生统一多模态模型。*arXiv* 预印本 *arXiv:2506.15564*，2025。

Xin, Y.、Qin, Q.、Luo, S.、Zhu, K.、Yan, J.、Tai, Y.、Lei, J.、Cao, Y.、Wang, K.、Wang, Y. 等。Lumina-dimoo：一种用于多模态生成与理解的全模态扩散大型语言模型。*arXiv* 预印本 *arXiv:2510.06308*，2025。

Yang, L.、Tian, Y.、Li, B.、Zhang, X.、Shen, K.、Tong, Y. 和 Wang, M. Mmada：多模态大型扩散语言模型。*arXiv* 预印本 *arXiv:2505.15809*，2025a。

Yang, Y.、Wang, C.、Wang, S.、Wen, Z.、Qi, B.、Xu, H. 和 Zhang, L. 具备原生可变生成长度的扩散大语言模型：让 [eos] 引领方向。*arXiv* 预印本 *arXiv:2510.24605*，2025b。

Ye, J.、Xie, Z.、Zheng, L.、Gao, J.、Wu, Z.、Jiang, X.、Li, Z.和Kong, L. Dream 7b：扩散大语言模型。*arXiv*预印本 *arXiv:2508.15487*，2025。

You, Z.、Nie, S.、Zhang, X.、Hu, J.、Zhou, J.、Lu, Z.、Wen, J.-R.和Li, C. LLaDA-V：采用视觉指令微调的大型语言扩散模型。*arXiv*预印本 *arXiv:2505.16933*，2025a。

You, Z.、Ou, J.、Zhang, X.、Hu, J.、Zhou, J.和Li, C. 高效的掩码图像生成模型。*arXiv*预印本 *arXiv:2503.07197*，2025b。

Yu, R.、Ma, X.和Wang, X. Dimple：采用并行解码的离散扩散多模态大语言模型。*arXiv*预印本 *arXiv:2505.16990*，2025。

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 等。Mmmu：面向专家级通用人工智能的大规模跨学科多模态理解与推理基准。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9556–9567 页，2024。

Zhai, X., Mustafa, B., Kolesnikov, A. 和 Beyer, L. 用于语言图像预训练的 Sig- moid 损失。载于*IEEE/CVF*计算机视觉国际会议论文集，第 11975–11986 页，2023。

Zhang, H., Duan, Z., Wang, X., Zhao, Y., Lu, W., Di, Z., Xu, Y., Chen, Y. 和 Zhang, Y. Nexus-gen：用于图像理解、生成和编辑的统一模型。*arXiv* 预印本 *arXiv:2504.21356*，2025a。Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Qiao, Y., 等。Math- verse：你的多模态大语言模型真的能看懂视觉数学问题中的图表吗？载于欧洲计算机视觉会议，第 169–186 页。施普林格，2024。

Zhang, Y., Ni, B., Chen, X.-S., Zhang, H.-R., Rao, Y., Peng, H., Lu, Q., Hu, H., Guo, M.-H., and Hu, S.- M. Bee：用于解锁先进的完全开放多模态大语言模型的高质量语料库和全栈套件。*arXiv*预印本 *arXiv:2510.13795*，2025b。

Zhao, S., Gupta, D., Zheng, Q., and Grover, A. d1：通过强化学习扩展扩散大语言模型的推理能力。*arXiv*预印本 *arXiv:2504.12216*，2025。

Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., and Levy, O. Transfusion：使用一个多模态模型预测下一个词元并扩散图像。*arXiv*预印本 *arXiv:2408.11039*，2024。

Zhou, J., Chen, H., Zhao, S., Kang, J., Li, J., Wang, E., Guo, Y., Sun, H., Wang, H., Kong, A., 等。Diffa：大型语言扩散模型能够聆听并理解。*arXiv* 预印本 *arXiv:2507.18452*，2025。

Zhu, F., Wang, R., Nie, S., Zhang, X., Wu, C., Hu, J., Zhou, J., Chen, J., Lin, Y., Wen, J.-R., 等。Llada 1.5：面向大型语言扩散模型的方差缩减偏好优化。*arXiv* 预印本 *arXiv:2505.19223*，2025a。

Zhu, F., You, Z., Xing, Y., Huang, Z., Liu, L., Zhuang, Y., Lu, G., Wang, K., Wang, X., Wei, L., 等。Llada-moe：一种稀疏 MoE 扩散语言模型。*arXiv* 预印本 *arXiv:2509.24389*，2025b。

Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., 和 Li, H. 从反思到完美：通过反思调优扩展文本到图像扩散模型的推理时优化。载于*IEEE/CVF*国际计算机视觉会议论文集，第 15329–15339 页，2025。

A. 相关工作

扩散大语言模型。近年来，扩散大语言模型（dLLMs）（Nie 等，2025；Ye 等，2025；Zhu 等，2025b）基于掩码扩散模型（MDMs）（Ou 等，2024；Lou 等，2024；Shi 等，2024；Sahoo 等，2024；You 等，2025b）发展而来，而掩码扩散模型是离散扩散模型（Sohl-Dickstein 等，2015；Hoogeboom 等，2021）的一种特殊形式。通过大规模预训练和监督微调，这些模型已达到与 LLaMA3 等强大的自回归模型相当的性能。这表明，dLLMs 作为自回归模型的替代方案具有实际应用价值和巨大潜力。除了文本生成之外，dLLMs 在多个领域也取得了显著进展，包括多模态理解（You 等，2025a；Yu 等，2025；Li 等，2025b）、音频理解（Zhou 等，2025）、强化学习（Zhu 等，2025a；Zhao 等，2025；Ou 等，2025）以及视觉-语言-动作任务（Wen 等，2025b；2025a；Liang 等，2025）。与我们的工作最相关的是统一多模态模型（Li 等，2025a；Xin 等人，2025；Yang 等，2025a）。然而，与依赖掩码扩散模型的这些方法不同，我们采用连续扩散模型进行图像生成，从而避免图像离散化造成的信息损失。

统一多模态模型最初以自回归架构为主（Team，2024a；Wang 等，2024），通过统一的下一词元预测，利用其与大语言模型在结构上的一致性。随后，混合方法将自回归文本处理与基于扩散的图像生成相结合（Zhou 等，2024；Xie 等，2024；Tong 等，2025；Liao 等，2025；Deng 等，2025）。与此同时，扩散大语言模型（dLLMs）（Nie 等，2025；Ye 等，2025；Zhu 等，2025b）通过大规模预训练和监督微调，已达到可与强大的自回归模型相媲美的性能。在这一成功的基础上，近期基于 dLLM 的统一多模态模型（Li 等人，2025a；Xin 等人，2025；Yang 等，2025a；Swerdlow 等，2025）采用掩码扩散机制来实现多模态理解与生成。

B. 实验细节

B.1. 训练阶段与数据设置

训练 LLaDA-o 的训练先经过一个初步对齐阶段，随后进入三个渐进式阶段 .

投影器对齐。 在正式训练之前，我们重点将视觉表征与理解专家对齐。我们使用来自 Infinity-MM（Gu, 2024）的阶段一数据训练多层感知机投影器，同时冻结所有其他组件。

阶段一：基础设置。在此阶段，我们使用大规模图像理解和生成数据建立基线能力。我们将图像生成的分辨率限制为 512，并且不对多模态理解应用自适应长度增强。对于文本，我们使用内部的 1000万条监督微调数据集。对于多模态理解，我们使用 Infinity-MM（Gu 等人，2024）的阶段 2–4 数据、MAmmoTH-VL-Instruct-12M（Guo 等人，2025）、LLaV A-OneVision-1.5（An 等人，2025）和 FineVision（An 等人，2025）。对于文本到图像生成，我们将图像描述数据（来自 Infinity-MM（Gu 等人，2024）和 LLaVA-OneVision-1.5（An 等人，2025））与生成数据集结合使用，包括 PD12M（Meyer 等人，2024）、Text-to-Image-2M（jackyhate，2024）、BLIP3o-Pretrain（Chen 等人，2025b）、Nexus-Gen（Zhang 等人，2025a）、FLUX-Reason-6M（Fang 等人，2025），以及来自 Qwen-Image（Wu 等人，2025b）的合成数据。

阶段二：高分辨率与推理。 我们进一步加入多模态推理数据，并将生成分辨率提高至 1024。自适应长度增强 对于理解任务仍保持禁用。数据方面，我们将 理解 数据来源切换为 Honey-Data-15M（Zhang 等人，2025b）。对于生成，我们通过移除图像说明和 PD12M（Meyer 等人，2024）来精简数据集，同时提高其余高质量数据的采样比例。此外，我们引入了来自 X2Edit（Ma 等人，2025）和 OmniGen2（Wu 等人，2025c）的交错多模态数据，以及来自 ShareGPT-4o-Image（Chen 等人，2025a）和 OpenGPT-4o-Image（Chen等人，2025d）的编辑数据。文本数据保持不变。

阶段三：可变长度优化。 在最后阶段，我们联合采用 自适应长度增强，为理解专家启用可变长度生成，并使用高质量数据微调模型。对于理解，我们引入 Honey-Data-1M，并保留 Honey-Data-15M（Zhang 等人，2025b），但降低其占比。对于生成，我们加入 ShareGPT-4o-Image（Chen等人，2025a）、OpenGPT-4o-Image（Chen等人，2025d）、BLIP3o-60k（Chen等人，2025b）和 GenRef（Zhuo 等人，2025）等优质数据集，同时降低阶段 2 生成数据的占比。对于交错数据，我们移除 X2Edit（Ma 等人，2025），其余部分则与阶段 2 保持一致。

B.2. 多轮对话数据与交错多模态数据设置

对于多轮对话，LLaDA（Nie 等人，2025）会随机选择一轮进行训练。它将此前的对话历史（包括提示和响应）拼接起来，作为该轮的输入上下文，并仅对选定轮次的响应计算损失。LaViDa（Li 等人，2025b）和 Dimple（Yu 等人，2025）也采用了这一设置。相比之下，LLaDA-V（You 等人，2025a）会对对话中每一轮的响应计算损失。在 LLaDA-o 中，我们沿用 LLaDA-V 的策略，因为该策略表现出色，且在 LLaDA-V 中展现了数据可扩展性。

对于交错多模态数据，我们仅将文本作为图像生成的条件（即提示词）。因此，我们仅针对每轮图像词元计算连续扩散损失，并屏蔽文本词元上的损失，以避免图像目标干扰文本表征。

B.3. 部分生成图像的提示词

表 7 列出了用于生成图 1所示样本的文本提示词。表中的图像编号对应图像的空间排列：编号 1–5 表示顶排（从左至右），编号 6–10 表示底排。

表 7. **生成样本的提示词**。 图像编号对应图 1中图像的排列顺序，即从左至右、从上至下。

图像 ID	提示词
1	明亮场景，俯视图，古城，奇幻，绚丽光线，镜面倒影，高细节，广角 lens.
2	一只逼真的柯基犬安静地坐在传统中式茶馆里，茶杯升起的蒸汽形成一条小龙，晨光，细致的毛皮。
3	一只身穿风衣的侦探猫站在雨巷中，赛博朋克霓虹招牌倒映在水洼里，黑色电影风格，体积雾，戏剧性阴影，电影镜头。
4	一名宇航员骑着马在月球上，梵高风格的油画。
5	复古灯泡里的微型图书馆，温暖舒适的灯光。
6	一只浣熊背着小背包，用地图在雾蒙蒙的山间小径上辨路，童话书插画 ，柔和色彩.
7	一座完全由折叠的折纸构成的复杂未来城市景观，柔和温暖的纸张质感，次表面散射，纸质建筑内部透出戏剧性照明，移轴摄影。
8	黑暗森林中的透明玻璃钢琴，钢琴内部生长着发光的生物发光蘑菇，空气中飘浮着魔法尘埃，奇幻插画，细节精致。
9	一辆完全由针织羊毛制成的红色跑车，柔和的影棚灯光。
10	未来机器人的肖像，列奥纳多·达·芬奇风格的铅笔素描。

B.4. 计算成本

表 8 详细说明了 LLaDA-o 每个训练阶段所需的计算资源和时间。主要训练阶段（阶段 1 和阶段 2）在 256 个 NVIDIA H800 GPU 上进行，占据了计算成本的绝大部分。最终的精炼阶段（阶段 3）使用 64 个 NVIDIA A100 GPU 完成。

B.5. LLaDA-o 的文本生成过程

在推理过程中，我们采用分块采样过程（Arriola 等人，2025）。如算法 2所示，我们首先缓存固定前缀（图像和提示词）。随后，通过追加一个长度为 L、以掩码标记初始化的块来扩展序列。在每个块内，我们进行迭代去噪：接受预测置信度超过阈值 τ 的词元，其余词元则继续保持掩码状态，以待下一步处理。如果检测到 ⌈序列结束符⌉ 词元，我们便截断序列并终止；否则，使用已完成的块更新缓存，并继续处理下一个掩码块。

表 *8.* **计算资源和训练成本**。 该表详细列出了每个训练阶段使用的图形处理器硬件及消耗的图形处理器总时数。

训练阶段	GPUs	图形处理器小时数
阶段 1	256 台 H800	55,296
阶段 2	256 H800	30,720
阶段 3	64 A100	1,536

算法 2 LLaDA-o 的文本生成过程。

输入： 参数 θ；图像 $\mathcal{V}$；提示词 $\mathcal{P}$；块大小 L；阈值 τ。
输出： 生成序列$\mathbf{y}$。
$\mathcal{C} \leftarrow$ 编码$(\mathcal{V}, \mathcal{P};\ \theta)$；$\mathbf{y} \leftarrow \emptyset$ {初始上下文编码}
循环
  $\mathbf{b} \leftarrow [\text{掩码}]^L$； $\mathcal{M} \leftarrow \{1, \ldots, L\}$ {初始化块和掩码索引}
  当 $\mathcal{M} \neq \emptyset$ **执行**
    $\mathbf{p} \leftarrow f_\theta(\mathbf{b}\ \mathcal{C})$ {使用参数进行前向传播 θ}
    对于 $i \in \mathcal{M}$**执行**
      如果$\max(\mathbf{p}_i) > \tau$ **则**
        $\mathbf{b}_i \leftarrow \arg\max(\mathbf{p}_i)$ {保留高置信度词元}
        $\mathcal{M} \leftarrow \mathcal{M} \setminus \{i\}$
      结束如果
    结束循环
  结束 while 循环
  如果 序列结束符 $\in \mathbf{b}$ **则**
    返回 $\mathbf{y} \parallel$ Trunc(**b**,序列结束符)
  结束 if
  $\mathbf{y} \leftarrow \mathbf{y} \parallel \mathbf{b}$；$\mathcal{C} \leftarrow$ 更新缓存$(\mathcal{C}, \mathbf{b})$
结束循环

C. 补充结果

C.1. 额外生成的图像

我们在图 5中提供了更多由 LLaDA-o 生成的定性示例。这些样例进一步展示了该模型生成高质量图像的能力：图像在语义上与用户提示保持一致，同时兼具高保真度和多样性。

C.2. 不匹配的块长度下与 LLaDA-V 的定性比较

本节展示定性样例，以说明 LLaDA-o 的可变长度文本生成能力。如表 9 所示，我们将 LLaDA-o 与最先进的 LLaDA-V 进行比较，考察提示词的语义要求与预定义生成块长度相冲突的场景。

首先，如表 9 顶部一行所示，当用户请求进行简单的文本提取，却提供了较长的块长度（即：$L = 64$）时，LLaDA-V 往往会用冗余内容填满整个窗口。相比之下，LLaDA-o 能够准确提取文本并正确终止，遵循用户意图。其次，如底部一行所示，当用户请求详细的图像描述，却分配了较短的块长度（即： $L = 16$）时，LLaDA-V 受到固定窗口的限制，导致响应过于简短。相反，如果在当前块中未检测到序列结束符（EOS），LLaDA-o 会自动追加额外的块来扩展生成，并持续生成，直至完成。

C.3. 块长度对变长生成的影响

本节展示定性样例，以说明 LLaDA-o 的可变长度生成行为。如表 10所示，我们保持图像和提示词不变，同时改变预定义生成块长度 $L \in \{16, 32, 64, 128\}$。



一位巨大的宇航员坐在金门大桥上钓鱼，云彩环绕，氛围如梦似幻，复古旅行海报风格，粉彩色，哑光绘画。



一只可爱的小机器人拿着红色气球，定格黏土动画风格，黏土上可见指纹，影棚灯光，景深，橡皮泥质感。



赛博朋克风格的《神奈川冲浪里》，海浪上霓虹灯闪耀，背景中是未来感的东京天际线，传统木版画质感与数字故障艺术相融合。



一只由透明玻璃制成的柯基犬彩色照片，它站在约塞米蒂国家公园的河流岸边。



一只背着迪斯科球壳的蜗牛，微距摄影，写实风格。



一匹由飞溅的蓝色水花构成的奔马，凝固动作摄影。



一个透明的玻璃汉堡包，内部霓虹灯闪耀，背景昏暗。



一朵由彩色烟雾构成的盛开玫瑰，黑色背景。



一台由透明玻璃制成的打字机，里面困着微小而鲜活的萤火虫，它们就是“字母”；打字机放在作家的木桌上，散发柔和的光芒。



一只章鱼在爵士俱乐部里演奏架子鼓，电影感灯光。



蜘蛛网上一滴露珠的微距摄影，水滴内部映出整个星系，极致特写，细节丰富，背景带有散景，氛围梦幻。



一匹由飞溅的液态水构成的雄伟骏马，在火海上飞奔，高速摄影，定格瞬间，水滴晶莹剔透，电影感灯光，蓝橙对比强烈。

图*5*。 **更多生成样本**。 我们展示了由 LLaDA-o 生成的 12 张随机选取的图像。每个样本所用的提示词显示在对应图像下方。所有结果均在与主论文相同的设置下生成。

在这些不同设置下，生成的输出总体上保持一致，响应长度仅有轻微波动。这表明 LLaDA-o 并未受到特定块长度的严格限制，并能够动态调整

有效的输出 长度，依据用户提示和图像内 容。

表 *9.* **展示可变长度生成的定性示例。上：** 在块长度较长的设置下（$L = 64$），LLaDA-V 会针对简单提示生成冗余内容，而 LLaDA-o 则能正确结束。**下：** 在块长度较短的设置下（$L = 16$），LLaDA-V 对详细提示的响应过于简短，而 LLaDA-o 则会根据用户提示动态调整输出长度。

User 图片上写了什么？直接告诉我答案。



LLaDA-V 图片中展示了一块灯光广告牌，上面用醒目、色彩鲜艳的字体写着“TIME FOR CHANGE”。广告牌背后是模糊的灯光，营造出节日或庆祝的氛围。牌上的文字是“TIME FOR CHANGE”，表达了行动号召或变革的信息。

LLaDA-o TIME FOR CHANGE

User 请详细描述这张图片。



LLaDA-V 图中展现了一幅宁静的景观：平静的湖泊被巍峨的山脉环绕。LLaDA-o 图中展现了一幅宁静的山地景观，可能是一处热门旅游目的地。画面主体是一片宽阔的湖泊，湖面呈鲜亮的绿色，倒映着周围的山脉。湖岸边设有木栅栏，为画面增添了质朴的韵味。栅栏外是一片草地，点缀着树木和一栋小建筑，可能是庇护所或游客中心。建筑掩映在树木之间，与自然环境融为一体。远处的山脉巍然耸立，山峰覆着薄雪，表明这里可能位于高海拔地区。上方的天空湛蓝晴朗，为整个画面增添了宁静的氛围。图中没有任何可见文字。

表*10*。 **展示可变长度生成的定性样本**。 我们展示了在不同块长度设置（$L \in \{16, 32, 64, 128\}$）下，LLaDA-o 针对同一图像和同一提示词生成的结果。尽管更改了 L，输出长度仍基本保持稳定，这表明 LLaDA-o 能够根据用户提示和图像动态调整输出长度。

User	请详细描述这张图像。  LLaDA-o 这张图像展现了一幅宁静的山地景观，很可能是一个热门的旅游目的地。场景的主体（块：16）是一片大型湖泊，湖面呈鲜艳的绿色，倒映着周围的山脉。湖边设有木栅栏，为场景增添了质朴的魅力。栅栏之外是一片草地，其间点缀着树木和一座小建筑，可能是庇护所或游客中心。这座建筑掩映在树木之间，与周围的自然环境融为一体。远处的山脉巍峨耸立，山峰覆盖着积雪，暗示这里位于高海拔地区。上方的天空湛蓝而澄澈，为整个场景增添了宁静的氛围。图像中没有任何可见的文字。
LLaDA-o (块:32)	这幅图描绘了一处宁静而如画的自然景观，湖泊平静，四周环绕着崎岖的 山脉和葱郁的绿意。湖泊呈鲜艳的绿松石色，倒映着上方晴朗的蓝天。一道 木栅栏沿着湖边延伸，将草地与水面隔开。栅栏简朴而质朴， 由垂直的木柱和横栏组成。前景中有一片维护良好的 有木栅栏的草地。草地青翠茂盛，呈现出健康自然的环境。 草地间点缀着几朵黄色花朵，为景色增添了一抹色彩。整体 图像氛围宁静而宜人，仿佛是放松身心、尽情欣赏 自然之美的绝佳去处。
LLaDA-o （块：64）	图中展现了一幅宁静的自然景观：一片清澈的小湖泊，四周环绕着葱郁的绿草地和 高大的常绿树。湖泊坐落在山谷中，背景处耸立着崎岖的岩石山脉。 天空晴朗湛蓝，显示这是一个阳光明媚的日子。前景中有一道木栅栏沿着 草地边缘延伸，将草地与湖泊隔开。右侧可见一座小型木制建筑，可能是小木屋或一栋 小房屋，部分被高大的树木遮挡。整个景色 宁静而如诗如画。
LLaDA-o （块：128）	这幅图描绘了一幅宁静的山地景观：一座清澈的小湖泊静卧在山谷中。 湖泊四周环绕着茂盛的绿草和树木，与 背景中的岩石山脉形成鲜明对比。崎岖的山峰巍然耸立，与晴朗的蓝天相映。 画面采用低角度视角，仰望群山，更添宏伟之感 并为这一场景增添雄伟之感。画面的整体氛围平和而宁静，邀请观者 沉浸于这片山地天堂的自然美景之中。