

# 视觉自回归建模：通过下一代预测进行可扩展的图像生成

Keyu Tian<sup>1,2</sup>, Yi Jiang<sup>2,†</sup>, Zehuan Yuan<sup>2,\*</sup>, Bingyue Peng<sup>2</sup>, Liwei Wang<sup>1,\*</sup>

<sup>1</sup>Peking University      <sup>2</sup>Bytedance Inc

keyutian@stu.pku.edu.cn, jiangyi.enjoy@bytedance.com,  
yuanzehuan@bytedance.com, bingyue.peng@bytedance.com, wanglw@pku.edu.cn

尝试探索我们的在线演示：<https://var.visionCodes> 和模型：<https://github.com/FoundationVision/VAR>

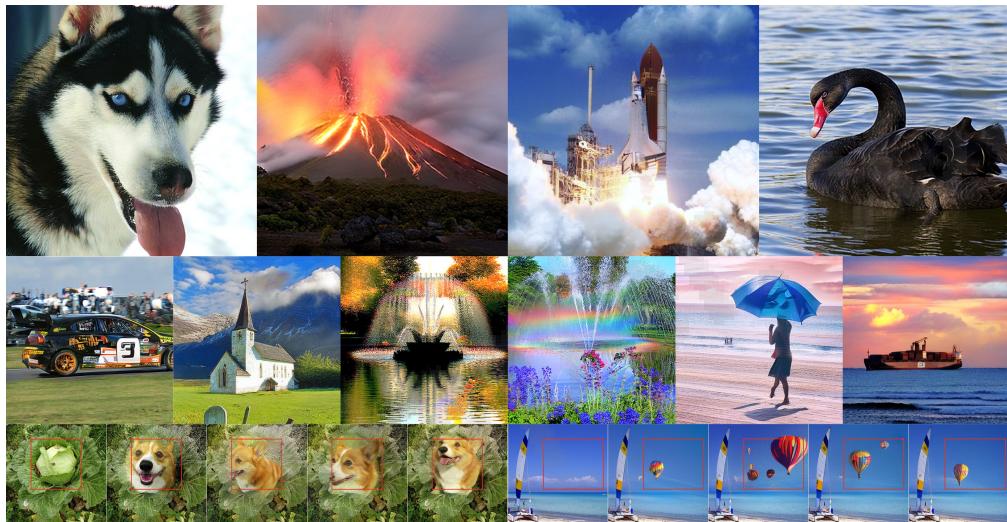


图 1：从在 ImageNet 上训练的视觉自回归 (VAR) 转换器生成的样本。我们显示  $512 \times 512$  样本（上）、 $256 \times 256$  样本（中）和零样本图像编辑结果（下）。

## Abstract

我们提出了视觉自回归建模 (VAR)，这是一种新一代范式，它将图像的自回归学习重新定义为从粗到细的“下一个尺度预测”或“下一个分辨率预测”，与标准光栅扫描的“下一个标记预测”不同”。这种简单、直观的方法允许自回归 (AR) 变压器快速学习视觉分布并且可以很好地泛化：VAR首次使GPT式AR模型在图像生成方面超越了扩散变压器。在 ImageNet  $256 \times 256$  基准上，VAR 通过将 Fréchet 起始距离 (FID) 从 18.65 提高到 1.73，起始分数 (IS) 从 80.4 提高到 350.2，显着提高了 AR 基线，推理速度提高了 20 倍。经验还验证了 VAR 在图像质量、推理速度、数据效率和可扩展性等多个维度上均优于 Diffusion Transformer (DiT)。放大 VAR 模型表现出与法学硕士中观察到的清晰的幂律缩放定律，线性相关系数接近 -0.998 作为确凿证据。VAR进一步展示了下游任务中的零样本泛化能力，包括图像内画、外画和编辑。这些结果表明 VAR 最初模拟了 LLM 的两个重要属性：缩放定律和零样本泛化。我们已经发布了所有模型和代码，以推动AR/VAR模型在视觉生成和统一学习方面的探索。

\* 通讯作者：wanglw@pku.edu.cn、yuanzehuan@bytedance.com； †：项目负责人

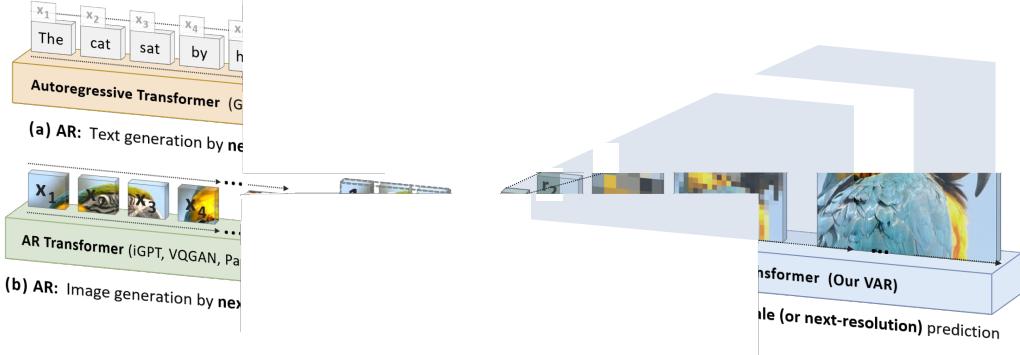


图2: 标准自回归模型(AR)与我们提出的视觉自回归模型(VAR)。

(a) AR应用于语言: 从左到右、逐字顺序生成文本标记; (b) AR应用于图像: 以光栅扫描顺序从左到右、从上到下顺序生成视觉标记; (c) 图像的VAR: 多尺度标记图是从粗到细尺度(从低到高分辨率)自回归生成的, 每个尺度内并行标记生成。VAR需要多尺度VQVAE才能发挥作用。

## 1 简介

GPT系列[66, 67, 15, 63, 1]和更多自回归(AR)大语言模型(LLM)[22, 4, 39, 83, 84, 91, 79, 5, 80]的出现预示着人工智能领域的新的纪元。这些模型在通用性和多功能性方面表现出了有前途的智能, 尽管存在幻觉等问题[40], 仍然被认为向通用人工智能(AGI)迈出了坚实的一步。这些模型的核心是自我监督学习策略——预测序列中的下一个标记, 这是一种简单而深刻的方法。对这些大型AR模型成功的研究强调了它们的可扩展性和泛化性: 前者, 如缩放定律[44, 36]所示, 使我们能够从较小的模型中预测大型模型的性能, 从而指导更好的资源分配, 而后者, 如零样本和少样本学习[67, 15]证明了这一点, 强调了无监督训练模型对各种看不见的任务的适应性。这些特性揭示了AR模型从大量未标记数据中学习的潜力, 概括了“AGI”的本质。

与此同时, 计算机视觉领域一直在努力开发大型自回归或世界模型[59, 58, 6], 旨在模拟其令人印象深刻的可扩展性和泛化性。VQGAN和DALL-E[30, 68]及其后继者[69, 92, 51, 99]等开创性的努力展示了AR模型在图像生成方面的潜力。这些模型利用视觉标记器将连续图像离散化为2D标记网格, 然后将其展平为用于AR学习的1D序列(图2b), 反映了顺序语言建模的过程(图2a)。然而, 这些模型的标度规律仍未得到充分探索, 更令人沮丧的是, 它们的性能明显落后于扩散模型[64, 3, 52], 如图3所示。与法学硕士的显著成就相比, 计算机中自回归模型的威力视线似乎有些被锁定。

自回归建模需要定义数据的顺序。我们的工作重新考虑如何“排序”图像: 人类通常以分层方式感知或创建图像, 首先捕获全局结构, 然后捕获局部细节。这种多尺度、从粗到细的性质暗示了一种“秩序”对于图像。同样受到广泛的多尺度设计[55, 53, 82, 45]的启发, 我们将图像的自回归学习定义为图2(c)中的“下一个尺度预测”, 与图2(c)中传统的“下一个标记预测”不同。2(b)。我们的方法首先将图像编码为多尺度标记图。然后, 自回归过程从 $1 \times 1$ 令牌图开始, 并逐步扩展分辨率: 在每一步, 变换器都会根据所有先前的令牌图来预测下一个更高分辨率的令牌图。我们将此方法称为视觉自回归(VAR)建模。

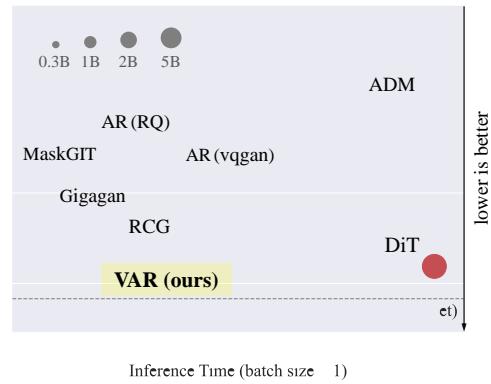


图3: 不同模型系列的缩放行为  
位于ImageNet  $256 \times 256$ 生成基准上。验证集的FID作为参考下限(1.78)。具有2B参数的VAR达到1.73的FID, 超过具有3B或7B参数的L-DiT。

VAR直接利用类似GPT-2的变压器架构[67]进行视觉自回归学习。在ImageNet  $256 \times 256$  基准上，VAR显着提高了其AR基线，实现了1.73的Fréchet起始距离（FID）和1.73的起始分数（IS）350.2，推理速度快20倍（详细信息请参见第6节）。值得注意的是，VAR在FID/IS、数据效率、推理速度和可扩展性方面超越了Diffusion Transformer（DiT）—Stable Diffusion 3.0和SORA [29, 14]等领先扩散系统的基础。VAR模型还表现出类似于法学硕士中所见的缩放定律。最后，我们展示了VAR在图像内画、外画和编辑等任务中的零样本泛化能力。总而言之，我们对社区的贡献包括：

1. 一种新的视觉生成框架，使用具有下一代预测的多尺度自回归范式，为计算机视觉的自回归算法设计提供了新的见解。
2. 对VAR模型的缩放定律和零样本泛化潜力进行实证验证，初步模拟了大型语言模型（LLM）的吸引人的特性。
3. 视觉自回归模型性能的突破，使得GPT式自回归方法在图像合成中首次超越强扩散模型<sup>2</sup>。
4. 全面的开源代码套件，包括VQ分词器和自回归模型训练管道，有助于推动视觉自回归学习的进步。

## 2 相关工作

### 2.1 大型自回归语言模型的性质

在自回归语言模型中发现并研究了尺度定律[44, 36]，它描述了模型（或数据集、计算等）的尺度与测试集上的交叉熵损失值之间的幂律关系。缩放定律使我们能够直接从较小的模型预测较大模型的性能[1]，从而指导更好的资源分配。更令人高兴的是，他们表明法学硕士的性能可以随着模型、数据和计算的增长而很好地扩展，并且永远不会饱和，这被认为是[15, 83, 84, 98, 91, 39]成功的关键因素。缩放定律带来的成功启发了视觉界探索更多类似的多模态理解和生成方法[54, 2, 89, 27, 96, 78, 21, 23, 42, 32, 33, 81, 88]。

零样本泛化。零样本泛化[73]是指模型（特别是大型语言模型）执行尚未明确训练的任务的能力。在计算机视觉领域，人们对基础模型CLIP [65]、SAM [49]、Dinov2 [62]的零样本和上下文学习能力越来越感兴趣。Painter [90]和LVM [6]等创新扩展了视觉提示器[41, 11]，以实现视觉中的上下文学习。

### 2.2 视觉生成

用于视觉生成的光栅扫描自回归模型需要将2D图像编码为1D标记序列。早期的努力[20, 85]已经展示了以标准的逐行光栅扫描方式生成RGB（或分组）像素的能力。[70]通过使用多个独立的可训练网络来重复进行超分辨率来扩展[85]。VQGAN [30]通过在VQVAE [86]的潜在空间中进行自回归学习来推进[20, 85]。它采用GPT-2仅解码器变压器以光栅扫描顺序生成标记，就像ViT [28]如何将2D图像序列化为1D补丁一样。VQVAE-2 [69]和RQ-Transformer [51]也遵循这种光栅扫描方式，但使用额外的尺度或堆叠代码。Parti [93]基于ViT-VQGAN [92]的架构，将变压器扩展到20B参数，并且在文本到图像合成中效果良好。

掩模预测模型。MaskGIT [17]采用类似于BERT [25, 10, 35]的VQ自动编码器和掩码预测变换器，通过贪婪算法生成VQ令牌。MagViT [94]使这种方法适用于视频，MagViT-2 [95]通过为图像和视频引入改进的VQVAE来增强[17, 94]。MUSE [16]进一步将MaskGIT扩展到3B参数。

扩散模型的进展集中在改进学习或采样[77, 76, 56, 57, 7]，指导[38, 61]，潜在学习[71]和架构[37, 64, 72, 31]。DiT和U-ViT [64, 8]用Transformer替换或集成了U-Net，并启发了最近的图像[19, 18]或视频合成系统[12, 34]，包括Stable Diffusion 3.0 [29]、SORA [14]、和维杜[9]。

<sup>2</sup>名为“语言模型击败扩散”的相关工作[95]属于BERT风格的掩模预测模型。

### 3 方法

#### 3.1 初步：通过下一个标记预测进行自回归建模

公式。考虑离散标记序列  $x = (x_1, x_2, \dots, x_T)$ , 其中  $x_t \in [V]$  是大小为  $V$  的词汇表中的整数。下一个标记自回归假设观察当前标记  $x_t$  的概率仅取决于其前缀  $(x_1, x_2, \dots, x_{t-1})$ 。这种单向标记依赖假设允许对序列  $x$  的可能性进行因式分解：

$$p(x_1, x_2, \dots, x_T) = \prod_{t=1}^T p(x_t | x_1, x_2, \dots, x_{t-1}). \quad (1)$$

训练自回归模型  $p(\theta)$  涉及在数据集上优化  $p(\theta)(x_t | x_1, x_2, \dots, x_{t-1})$ 。这被称为“下一个标记预测”，训练后的  $p(\theta)$  可以生成新序列。

货币化。图像本质上是二维连续信号。要通过下一个标记预测将自回归建模应用于图像，我们必须：1) 将图像标记为多个离散标记，2) 为单向建模定义标记的一维顺序。对于1)，通常使用量化自动编码器（例如[30]）将图像特征图  $f \in \mathbb{R}^{h \times w \times C}$  转换为离散标记  $q \in [V]^{h \times w}$ ：

$$f = \mathcal{E}(im), \quad q = \mathcal{Q}(f), \quad (2)$$

其中  $im$  表示原始图像， $\mathcal{E}(\cdot)$  表示编码器， $\mathcal{Q}(\cdot)$  表示量化器。量化器通常包括包含  $V$  个向量的可学习码本  $Z \in \mathbb{R}^{V \times C}$ 。量化过程  $q = \mathcal{Q}(f)$  将把每个特征向量  $f(i,j)$  映射到欧几里得意义上最接近的代码的代码索引  $q(i,j)$ ：

$$q^{(i,j)} = \left( \arg \min_{v \in [V]} \| \text{lookup}(Z, v) - f^{(i,j)} \|_2 \right) \in [V], \quad (3)$$

其中  $\text{lookup}(Z, v)$  表示获取码本  $Z$  中的第  $v$  个向量。为了训练量化自动编码器，每个  $q(i,j)$  查找  $Z$  以获得  $f$ ，即原始  $f$  的近似值。然后使用给定  $f$  的解码器  $\mathcal{D}(\cdot)$  重建新图像  $\hat{im}$ ，并最小化复合损失  $L$ ：

$$f = \text{查找}(Z, q), \hat{im} = \mathcal{D}(f), \quad (4)$$

$$\mathcal{L} = \|im - \hat{im}\|_2 + \|f - \hat{f}\|_2 + \lambda_P \mathcal{L}_P(im) + \lambda_G \mathcal{L}_G(\hat{im}), \quad (5)$$

其中  $\mathcal{L}_P(\cdot)$  是感知损失，如 LPIPS [97]， $\mathcal{L}_G(\cdot)$  是判别损失，如 StyleGAN 的判别器损失 [47]，而  $\lambda_P$ 、 $\lambda_G$  是损失权重。一旦自动编码器  $\{\mathcal{E}, \mathcal{Q}, \mathcal{D}\}$  得到充分训练，它将用于标记图像以用于单向自回归模型的后续训练。

$q \in [V]^{h \times w}$  中的图像标记排列在 2D 网格中。与具有固有的从左到右排序的自然语言句子不同，必须为单向自回归学习明确定义图像标记的顺序。以前的 AR 方法 [30, 92, 51] 使用一些策略（例如行主栅格扫描、螺旋或 z 曲线）将  $q$  的 2D 网格展平为 1D 序列  $x = (x_1, \dots, x_{h \times w})$  一旦展平，他们就可以从数据集中提取一组序列  $x$ ，然后训练自回归模型，通过下一个标记预测来最大化 (1) 中的可能性。

关于普通自回归模型的弱点的讨论。上述标记化和平滑方法可以实现图像上的下一个标记自回归学习，但引入了几个问题：

- 1) 违反数学前提。在量化自动编码器 (VQVAE) 中，编码器通常会生成图像特征图  $f$ ，其中所有  $i, j$  都具有相互依赖的特征向量  $f(i,j)$ 。因此，在量化和平滑之后，令牌序列  $(x_1, x_2, \dots, x_{h \times w})$  保留双向相关性。这与自回归模型的单向依赖性假设相矛盾，该假设规定每个标记  $x_t$  应该仅依赖于其前缀  $(x_1, x_2, \dots, x_{t-1})$ 。2) 无法执行一些零样本泛化。与问题 1) 类似，图像自回归建模的单向性质限制了它们在需要双向推理的任务中的通用性。例如，它无法在给定底部的情况下预测图像的顶部。3) 结构退化。扁平化破坏了图像特征图中固有的空间局部性。例如，令牌  $q(i,j)$  及其 4 个直接邻居  $q(i \pm 1, j)$ 、 $q(i, j \pm 1)$  由于邻近而紧密相关。这种空间关系在线性序列中受到损害，其中单向约束减少了这些相关性。

4) 效率低下。使用传统的自注意力变换器生成图像令牌序列  $x = (x_1, x_2, \dots, x_{n \times n})$  会产生  $O(n^2)$  自回归步骤和  $O(n^6)$  计算成本。

问题 2) 和 3) 很明显 (参见上面的示例)。关于问题 1)，我们在附录 C 中提供了经验证据。问题 3) 的证明在附录 D 中详细说明。这些理论和实践限制要求在图像生成的背景下重新思考自回归模型。

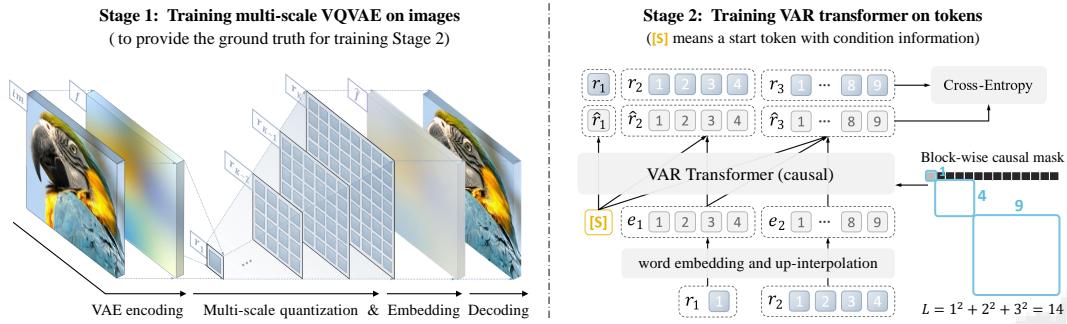


图 4: VAR 涉及两个独立的训练阶段。第 1 阶段: 多尺度 VQ 自动编码器编码

将图像转换为  $K$  个令牌映射  $R = (r_1, r_2, \dots, r_K)$  并通过复合损失 (5) 进行训练。有关“多尺度量化”和“嵌入”的详细信息，请查看算法 1 和 2。第 2 阶段: 通过下一尺度预测训练 VAR 变换器 (6): 它需要  $([s], r_1, r_2, \dots, r_{K-1})$  作为预测  $(r_1, r_2, r_3, \dots, r_K)$  的输入。训练中使用注意力掩码来确保每个  $r_k$  只能关注  $r \leq k$ 。使用标准交叉熵损失。

### 3.2 通过下一尺度预测的视觉自回归建模

重新制定。我们通过从“下一个令牌预测”策略转变为“下一个尺度预测”策略，重新概念化了图像的自回归模型。这里，自回归单元是整个 tokenmap，而不是单个 token。我们首先将特征映射  $f \in \mathbb{R}^{h \times w \times C}$  量化为  $K$  个多尺度标记映射  $(r_1, r_2, \dots, r_K)$ ，每个映射的分辨率越来越高  $h_k \times w_k$ ，最终得到  $r_K$  matches 原始特征图的分辨率  $h \times w$ 。自回归可能性的公式为:

$$p(r_1, r_2, \dots, r_K) = \prod_{k=1}^K p(r_k | r_1, r_2, \dots, r_{k-1}), \quad (6)$$

其中每个自回归单元  $r_k \in [V]$   $h_k \times w_k$  是包含  $h_k \times w_k$  个标记的  $k$  尺度标记图，序列表  $(r_1, r_2, \dots, r_{k-1})$  作为“前缀”对于  $r_k$ 。在第  $k$  个自回归步骤中， $r_k$  中的  $h_k \times w_k$  标记上的所有分布将并行生成，以  $r_k$  的前缀和相关的第  $k$  个位置嵌入图为条件。这种“下一尺度预测”方法就是我们定义的视觉自回归模型 (VAR)，如图 4 右侧所示。请注意，在 VAR 训练中，使用了逐块因果注意掩模来确保每个  $r_k$  只能注意它的前缀  $r \leq k$ 。推理时可以使用 kv-caching，不需要 mask。

讨论。VAR 解决了前面提到的三个问题，如下：1) 如果我们约束每个  $r_k$  仅依赖于其前缀，即得到  $r_k$  的过程仅与  $r \leq k$  相关，则满足数学前提。这种约束是可以接受的，因为  $s$ ，它符合自然的、从粗到细的进展特征，如人类视觉感知和艺术绘画（正如我们在第 1 节中讨论的）。下面的代币化中提供了更多详细信息。

2) 空间局部性被保留，因为 (i) VAR 中没有展平操作，并且 (ii) 每个  $r_k$  中的标记完全相关。多尺度的设计进一步强化了空间结构。

3) 生成具有  $n \times n$  潜在图像的复杂度显着降低至  $O(n^4)$ ，证明见附录。这种效率增益来自于每个  $r_k$  中的并行令牌生成。

代币化。我们开发了一种新的多尺度量化自动编码器，将图像编码为 VAR 学习所需的  $K$  多尺度离散标记图  $R = (r_1, r_2, \dots, r_K)$  (6)。我们采用与 VQGAN [30] 相同的架构，但具有修改后的多尺度量化层。对  $f$  或  $f$  进行残差设计的编码和解码过程在算法 1 和 2 中详细说明。我们凭经验发现这种类似于 [51] 的残差式设计可以比独立插值表现得更好。算法 1 表明每个  $r_k$  仅取决于其前缀  $(r_1, r_2, \dots, r_{k-1})$ 。

请注意，共享码本  $Z$  在所有尺度上使用，确保每个  $rk$  的标记属于相同的词汇表  $[V]$ 。为了解决将  $z_k$  放大到  $h_K \times w_K$  时的信息丢失问题，我们使用 Kextra 卷积层  $\{\phi_k\}_{k=1}^K$ 。将  $f$  下采样为  $h_K \times w_K$  后不使用卷积。

算法1：多尺度VQVAE编码		算法2：多尺度VQVAE重建	
1	输入：原始图像 $im$ ; 2 超参数：步 长 $K$ 、分辨率( $hk$ , $wk$ ) $Kk=1$ ; 3 $f =$ $E(im)$ , $R = []$ ; 4 for $k = 1, \dots, K$ do	1	输入：多尺度 token 映射 $R$ ; 2 超参数：步骤 $K$ 、分辨率( $hk$ , $wk$ ) $Kk=1$ ; 3 $f = 0$ ; 4 for $k = 1, \dots, K$ do
5	$r_k = Q(\text{interpolate}(f, hk, wk));$	5	$r_k = \text{queue\_pop}(R);$
6	$R = \text{queue\_push}(R, r_k);$	6	$z_k = \text{lookup}(Z, r_k);$
7	$z_k = \text{lookup}(Z, r_k);$	7	$z_k = \text{interpolate}(z_k, hk, wk);$
8	$z_k = \text{interpolate}(z_k, hk, wk);$	8	$\hat{f} = \hat{f} + \phi_k(z_k);$
	$f = f - \text{phik}(zk);$ 10 返回：多尺度 标记 $R$ ;	9	$\hat{im} = (\hat{f});$
		10	返回：重建图像 $\hat{im}$ ;
			$\hat{m};$

4 实施细节

VAR 分词器。如上所述，我们使用普通的 VQVAE 架构 [30] 和具有  $K$  个额外卷积（0.03M 额外参数）的多尺度量化方案。我们对  $V = 4096$  的所有尺度使用共享代码本。按照基线 [30]，我们的分词器也在 OpenImages [50] 上进行训练，复合损失 (5) 和空间下采样率为  $16 \times$ 。

无功变压器。我们主要关注 VAR 算法，因此我们保持简单的模型架构设计。我们采用类似于 GPT-2 和 VQ-GAN [67, 30] 的标准解码器变压器架构，具有自适应归一化 (AdaLN)，该架构在许多视觉生成模型中得到了广泛采用并被证明有效 [47, 48, 46, 75]、74、43、64、19]。对于类条件合成，我们使用类嵌入作为起始标记  $[s]$  以及 AdaLN 的条件。我们现在注意力可以稳定训练之前将查询和键标准化为单位向量。我们没有在大型语言模型中使用先进技术，例如旋转位置嵌入 (RoPE)、SwiGLU MLP 或 RMS Norm [83, 84]。我们的模型形状遵循像[44]这样的简单规则，即宽度  $w$ 、头数  $h$  和掉落率  $dr$  随深度  $d$  线性缩放，如下所示：

$$w = 64d, \quad h = d, \quad dr = 0.1 \cdot d/24. \quad (7)$$

因此，深度为  $d$  的 VAR 变压器的主要参数计数  $N$  由下式给出： $3:N(d) = d \cdot 4w^2$

(8)

所有模型都使用类似的设置进行训练：每 256 个批量大小  $10^{-4}$  的基本学习率，AdamW 优化器， $\beta_1 = 0.9$ ， $\beta_2 = 0.95$ ，衰减 = 0.05，批量大小从 768 到 1024，训练周期从 200 至 350（取决于型号尺寸）。秒中的评估。图 5 表明这种简单的模型设计能够很好地扩展和泛化。

5 实证结果

本节首先将 VAR 与其他图像生成模型系列进行比较。5.1.第 2 节介绍了对 VAR 模型的可扩展性和泛化性的评估。5.2 和附录 B。实施细节和消融研究参见附录 4 和附录 6。

## 5.1 最先进的图像生成

设置。我们在 ImageNet  $256 \times 256$  和  $512 \times 512$  条件生成基准上测试深度为 16、20、24 和 30 的 VAR 模型，并将它们与最先进的图像生成模型系列进行比较。在所有基于 VQVAE 的 AR 或 VAR 模型中，VQGAN [30] 和我们的 VQVAE 使用相同的架构（CNN）和训练数据（OpenImages [50]），而 ViT-VQGAN [92] 使用 ViT 自动编码器，并且它和 RQTransformer [51] 直接在 ImageNet 上训练 VQVAE。结果总结在表中。1 和选项卡。2。

3 由于资源限制，我们在  $512 \times 512$  合成中的所有注意力块上使用单个共享自适应层规范 (AdaLN)。在这种情况下，参数数量将减少到大约  $12dw^2 + 6w^2 \approx 49152d^3$ 。

表 1：类条件 ImageNet 256×256 上的生成模型系列比较。“↓”或“↑”表示

值越低或越高越好。指标包括 Fréchet 起始距离 (FID)、起始分数 (IS)、精度 (Pre) 和召回率 (rec)。“#Step”：生成图像所需的模型运行次数。报告与 VAR 相关的挂钟推理时间。带有后缀 “-re”的模型使用拒绝采样。†：取自 MaskGIT [17]。

Type	Model	FID↓	IS↑	Pre↑	Rec↑	#Para	#Step	Time
GAN	BigGAN [13]	6.95	224.5	<b>0.89</b>	0.38	112M	1	—
GAN	GigaGAN [43]	3.45	225.5	0.84	<b>0.61</b>	569M	1	—
GAN	StyleGAN-XL [75]	2.30	265.1	0.78	0.53	166M	1	0.3 [75]
Diff.	ADM [26]	10.94	101.0	0.69	0.63	554M	250	168 [75]
Diff.	CDM [37]	4.88	158.7	—	—	—	8100	—
Diff.	LDM-4-G [71]	3.60	247.7	—	—	400M	250	—
Diff.	DiT-L/2 [64]	5.02	167.2	0.75	0.57	458M	250	31
Diff.	DiT-XL/2 [64]	2.27	278.2	0.83	0.57	675M	250	45
Diff.	L-DiT-3B [3]	2.10	304.4	0.82	0.60	3.0B	250	>45
Diff.	L-DiT-7B [3]	2.28	316.2	0.83	0.58	7.0B	250	>45
Mask.	MaskGIT [17]	6.18	182.1	0.80	0.51	227M	8	0.5 [17]
Mask.	RCG (cond.) [52]	3.49	215.5	—	—	502M	20	1.9 [52]
AR	VQVAE-2 <sup>†</sup> [69]	31.11	~45	0.36	0.57	13.5B	5120	—
AR	VQGAN <sup>†</sup> [30]	18.65	80.4	0.78	0.26	227M	256	19 [17]
AR	VQGAN [30]	15.78	74.3	—	—	1.4B	256	24
AR	VQGAN-re [30]	5.20	280.3	—	—	1.4B	256	24
AR	ViTVQ [92]	4.17	175.1	—	—	1.7B	1024	>24
AR	ViTVQ-re [92]	3.04	227.4	—	—	1.7B	1024	>24
AR	RQTran. [51]	7.55	134.0	—	—	3.8B	68	21
VAR	VAR-d16	3.30	274.4	0.84	0.51	310M	10	0.4
VAR	VAR-d20	2.57	302.6	0.83	0.56	600M	10	0.5
VAR	VAR-d24	2.09	312.9	0.82	0.59	1.0B	10	0.6
VAR	VAR-d30	1.92	323.1	0.82	0.59	2.0B	10	1
VAR	VAR-d30-re (validation data)	<b>1.73</b>	<b>350.2</b>	0.82	0.60	2.0B	10	1

整体比较。与现有的生成方法（包括生成对抗网络（GAN）、扩散模型（Diff.）、BERT 式掩模预测模型（Mask.）和 GPT 式自回归模型（AR）相比，我们的视觉自回归（VAR）建立了新模型类。如表所示。1、VAR不仅实现了最佳的FID/IS，而且在图像生成方面表现出了惊人的速度。VAR还保持了不错的精确度和召回率，证实了其语义一致性。这些优势在 512×512 综合基准上同样适用，如表 1 所示。2.值得注意的是，VAR显着提升了传统 AR能力。据我们所知，这是自回归模型首次超越扩散变压器，这是由于 VAR 解决了第 3 节中讨论的 AR 限制而成为可能的里程碑。

效率比较。传统的自回归（AR）模型 [30,69,92,51] 受到高计算成本的影响，因为图像标记的数量与图像分辨率成二次方。 $n^2$  令牌的完全自回归生成需要  $O(n^2)$  解码迭代和  $O(n^6)$  总计算。相反，VAR 只需要  $O(\log(n))$  次迭代和  $O(n^4)$  总计算。表中报告的挂钟时间。图1还提供了经验证据，表明即使模型参数更多，VAR也比VQGAN和 ViT-VQGAN快约20倍，达到高效GAN模型的速度，只需1步即可生成图像。

表 2：ImageNet 512×512 条件生成。

†：引自 MaskGIT [17]。“-s”：由于资源限制，使用单个共享 AdaLN 层。

Type	Model	FID↓	IS↑	Time
GAN	BigGAN [13]	8.43	177.9	—
Diff.	ADM [26]	23.24	101.0	—
Diff.	DiT-XL/2 [64]	3.04	240.8	81
Mask.	MaskGIT [17]	7.32	156.0	0.5 <sup>†</sup>
AR	VQGAN [30]	26.52	66.8	25 <sup>†</sup>
VAR	VAR-d36-s	<b>2.63</b>	<b>303.2</b>	1

与流行的扩散变压器相比。VAR模型在多个维度上超越了最近流行的扩散模型Diffusion Transformer (DiT)，它是最新的Stable-Diffusion 3 [29]和SORA [14]的前身：1) 在图像生成多样性和质量方面

(FID 和 IS)，具有 2B 参数的 VAR 始终比 DiT-XL/2 [64]、L-DiT-3B 和 L-DiT-7B [3] 表现更好。VAR 还保持了相当的精确度和召回率。2) 对于推理速度，与 VAR 相比，DiT-XL/2 需要 45 倍的挂钟时间，而 3B 和 7B 模型 [3] 的成本会更高。3) VAR 被认为数据效率更高，因为与 DiT-XL/2 的 1400 个训练周期相比，它只需要 350 个训练周期。4) 对于可扩展性，图 3 和表 3 所示。图 1 表明 DiT 在超过 675M 个参数时仅获得边际增益甚至负增益。相比之下，VAR 的 FID 和 IS 持续改进，与第 2 节中的标度律研究一致。5.2. 这些结果表明 VAR 可能是比 DiT 等模型更高效且可扩展的图像生成模型。

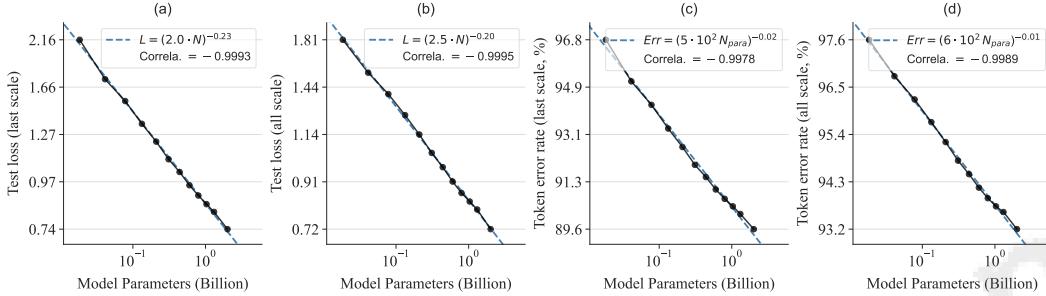


图 5: VAR 变压器尺寸  $N$  的缩放定律，以及幂律拟合（虚线）和方程（图例中）。  
小且接近于零的指数  $\alpha$  表明，当放大 VAR 变压器时，测试损耗  $L$  和令牌错误率  $Err$  都会平稳下降。轴均采用对数刻度。Pearson 相关系数接近 -0.998，表明  $\log(N)$  与  $\log(L)$  或  $\log(N)$  与  $\log(Err)$  之间存在很强的线性关系。

## 5.2 幂律缩放定律

背景。先前的研究 [44,36,39,1] 已经确定，扩大自回归 (AR) 大语言模型 (LLM) 会导致测试损失  $L$  的可预测下降。这种趋势与参数计数  $N$ 、训练标记  $T$  和最佳训练计算相关  $C \text{ min}$ ，遵循幂律：

$$L = (\beta \cdot X)^\alpha, \quad (9)$$

其中  $X$  可以是  $N$ 、 $T$  或  $C \text{ min}$  中的任意一个。指数  $\alpha$  反映幂律的平滑度， $L$  表示由不可约损失  $L \infty$  归一化的可约损失 [36] 对  $L$  和  $X$  的对数变换将揭示  $\log(L)$  和  $\log(X)$  之间的线性关系：

$$\log(L) = \alpha \log(X) + \alpha \log \beta. \quad (10)$$

一个吸引人的现象是，[44] 和 [36] 从未观察到  $X$  较高端与这些线性关系的偏差，尽管随着损失接近零，平坦化是不可避免的。

这些观察到的缩放定律 [44,36,39,1] 不仅验证了 LLM 的可扩展性，而且还可以作为 AR 建模的预测工具，有助于根据较小的模型评估较大 AR 模型的性能，从而节省资源使用大型模型性能预测。鉴于法学硕士带来的缩放定律的这些吸引人的特性，它们在计算机视觉中的复制因此引起了人们的极大兴趣。

设置缩放 VAR 模型。遵循 [44,36,39,1] 中的协议，我们检查我们的 VAR 模型是否符合类似的缩放定律。我们在每个 epoch 包含 1.28M 图像（或我们的 VQVAE 下的 870B 图像标记）的 ImageNet 训练集 [24] 上训练了 12 种不同大小（从 18M 到 2B 参数）的模型。对于不同规模的模型，训练跨越 200 到 350 个 epoch，最大代币数量达到 3050 亿个。下面我们将重点关注模型参数  $N$  的缩放定律和给定足够令牌计数  $T$  的最佳训练计算  $C \text{ min}$ 。

模型参数  $N$  的缩放法则。我们首先研究随着 VAR 模型规模的增加，测试损失的趋势。对于深度为  $d$  的 VAR 变压器，参数数量  $N(d) = 73728 d^3$  在 (8) 中指定。我们将  $d$  从 6 更改为 30，产生了 12 个模型，参数为 18.5M 到 2.0B。我们在包含 50,000 张图像的 ImageNet validation 集上评估了最终测试交叉熵损失  $L$  和标记预测错误率  $Err$  [24]。我们计算了最后一个尺度（在最后一个下一个尺度自回归步骤）以及全局平均值的  $L$  和  $Err$ 。结果绘制在图 5 中，其中我们

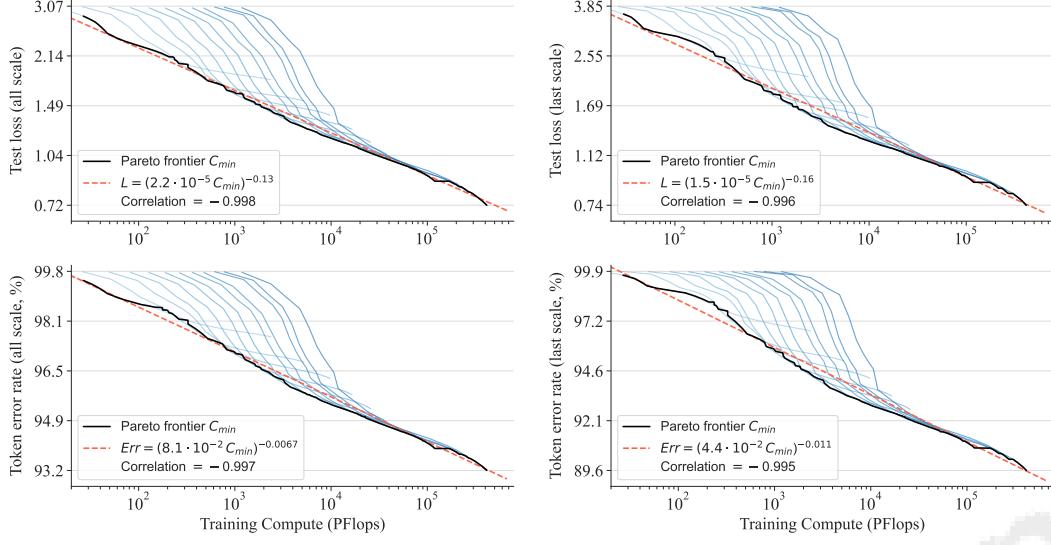


图 6：使用最佳训练计算  $C_{min}$  的缩放法则。线条颜色表示不同的模型尺寸。红色的虚线是与图例中的方程的幂律拟合。轴采用对数刻度。接近 -0.99 的 Pearson 系数表明  $\log(C_{min})$  与  $\log(L)$  或  $\log(C_{min})$  与  $\log(\text{Err})$  之间存在很强的线性关系。

观察到  $L$  作为  $N$  的函数的明显幂律缩放趋势，与 [44, 36, 39, 1] 一致。幂律缩放定律可以表示为：

$$L_{\text{last}} = (2.0 \cdot N)^{-0.23} \quad \text{and} \quad L_{\text{avg}} = (2.5 \cdot N)^{-0.20}. \quad (11)$$

虽然缩放定律主要是在测试损失上研究的，但我们也凭经验观察到令牌错误率 Err 的类似幂律趋势：

$$\text{Err}_{\text{last}} = (4.9 \cdot 10^2 N)^{-0.016} \quad \text{and} \quad \text{Err}_{\text{avg}} = (6.5 \cdot 10^2 N)^{-0.010}. \quad (12)$$

这些结果验证了 VAR 强大的可扩展性，通过扩大 VAR 变压器可以不断提高模型的测试性能。

具有最佳训练的缩放法则计算  $C_{min}$ 。然后，我们在增加训练计算  $C$  时检查 VAR 变压器的缩放行为。对于 12 个模型中的每一个，我们在 PFlops（每秒 10 15 次浮点运算）中引用的训练期间将测试损失  $L$  和令牌错误率 Err 跟踪为  $C$  的函数。结果如图 6 所示。在这里，我们绘制了  $L$  和 Err 的帕累托前沿，以突出显示达到特定损失或误差值所需的最佳训练计算  $C_{min}$ 。

$L$  和 Err 作为  $C_{min}$  函数的拟合幂律缩放定律为：

$$L_{\text{last}} = (2.2 \cdot 10^{-5} C_{min})^{-0.13} \quad (13)$$

$$L_{\text{avg}} = (1.5 \cdot 10^{-5} C_{min})^{-0.16}, \quad (14)$$

$$\text{最后误差} = (8.1 \cdot 10^{-2} C_{min})^{-0.0067} \quad (15)$$

$$\text{平均误差} = (4.4 \cdot 10^{-2} C_{min})^{-0.011}. \quad (16)$$

这些关系 (14, 16) 在  $C_{min}$  中保持了 6 个数量级，我们的发现与 [44, 36] 中的结果一致：当使用足够的数据进行训练时，较大的 VAR 变压器的计算效率更高，因为它们可以达到相同的效果计算量较少的性能水平。

## 6 消融研究

在本研究中，我们旨在验证我们提出的 VAR 框架的有效性和效率。结果如表 1 所示。3.

VAR 的有效性和效率。从 [17] 实现的普通 AR 变压器基线开始，我们用 VAR 替换其方法，并保持其他设置不变以获得第 2 行。

表3: VAR的消融研究。前两行比较在AR或AR下训练的GPT-2式Transformer

VAR算法没有任何附加功能。后续行显示了VAR增强的影响。“AdaLN”：自适应层范数。“CFG”：无分类器指导。“注意。Norm.”：在关注之前将q和k归一化为单位向量。“成本”：相对于基线的推理成本。“Delta”：FID降低至基线。

	Description	Para.	Model	AdaLN	Top- $k$	CFG	Cost	FID $\downarrow$	$\Delta$
1	AR [30]	227M	AR	✗	✗	✗	1	18.65	0.00
2	AR to VAR	207M	VAR-d16	✗	✗	✗	0.013	5.22	-13.43
3	+AdaLN	310M	VAR-d16	✓	✗	✗	0.016	4.95	-13.70
4	+Top- $k$	310M	VAR-d16	✓	900	✗	0.016	4.64	-14.01
5	+CFG	310M	VAR-d16	✓	900	1.5	0.022	3.60	-15.05
5	+Attn. Norm.	310M	VAR-d16	✓	900	1.5	0.022	3.30	-15.35
6	+Scale up	2.0B	VAR-d30	✓	900	1.5	0.052	1.73	-16.85

与AR模型相比，VAR仅用0.013倍的推理挂钟成本就实现了更好的FID(18.65 vs. 5.22)，这表明视觉AR模型的性能和效率得到了飞跃。

逐组件消融。我们进一步测试了VAR中的一些关键组件。通过用自适应层标准化(AdaLN)替换标准层标准化(LN)，VAR开始产生比基线更好的FID。通过使用与基线类似的top-k采样，VAR的FID进一步提高。通过使用比率为1.5的无分类器引导(CFG)并在注意之前将q和k归一化为单位向量，我们达到了3.30的FID比基线低15.35，推理速度仍快45倍。我们最终将VAR大小扩大到2.0B，并实现FID为1.73。这比基线FID好16.85。

## 7 局限性和未来的工作

在这项工作中，我们主要关注学习范式的设计，并保持VQVAE架构和训练与基线[30]保持不变，以更好地证明VAR框架的有效性。我们期望推进VQVAE分词器[99,60,95]作为增强自回归模型的另一种有前途的方法，这与我们的工作正交。我们相信在这些最新工作中通过先进的分词器或采样技术迭代VAR可以进一步提高VAR的性能或速度。

文本提示生成是我们研究的一个持续方向。鉴于我们的模型在本质上与现代法学硕士相似，因此它可以轻松地与现代法学硕士集成，通过编码器-解码器或上下文方式执行文本到图像的生成。

本工作中没有实现视频生成，但可以自然地扩展。通过将多尺度视频特征视为3D金字塔，我们可以制定类似的“3D下一尺度预测”来通过VAR生成视频。与SORA[14]等基于扩散的生成器相比，我们的方法在时间一致性或与LLM集成方面具有固有的优势，因此可以处理更长的时间依赖性。这使得VAR在视频生成领域具有竞争力，因为传统的AR模型由于其极高的计算复杂度和缓慢的推理速度，对于视频生成来说效率太低：用传统AR模型生成高分辨率视频变得非常昂贵，而VAR有能力解决这个问题。因此，我们预见到在视频生成领域利用VAR模型有着广阔的前景。

## 8 结论

我们引入了一种名为视觉自回归建模(VAR)的新视觉生成框架，它1)理论上解决了标准图像自回归(AR)模型固有的一些问题，2)使基于语言模型的AR模型在图像质量方面首先超越了强扩散模型，多样性、数据效率和推理速度。将VAR扩展到20亿个参数后，我们观察到测试性能与模型参数或训练计算之间存在明显的幂律关系，皮尔逊系数接近-0.998，表明性能预测的稳健框架。这些缩放定律和零样本任务泛化的可能性，作为法学硕士的标志，现已在我们的VAR变压器模型中得到初步验证。我们希望我们的发现和开源能够促进将自然语言处理领域的实质性成功更无缝地集成到计算机视觉中，最终促进强大的多模式智能的进步。

## 缩放效果的可视化

为了更好地理解 VAR 模型在放大时如何学习，我们比较了 4 个不同大小（深度 6、16、26、30）和 3 个不同训练阶段（20%、60%、100%）的 VAR 模型生成的一些  $256 \times 256$  样本。总训练令牌）如图 7 所示。为了保持内容一致，使用相同的随机种子和教师强制的初始令牌。观察到的视觉保真度和健全性的改进与缩放定律一致，因为较大的变换器被认为能够学习更复杂和更细粒度的图像分布。

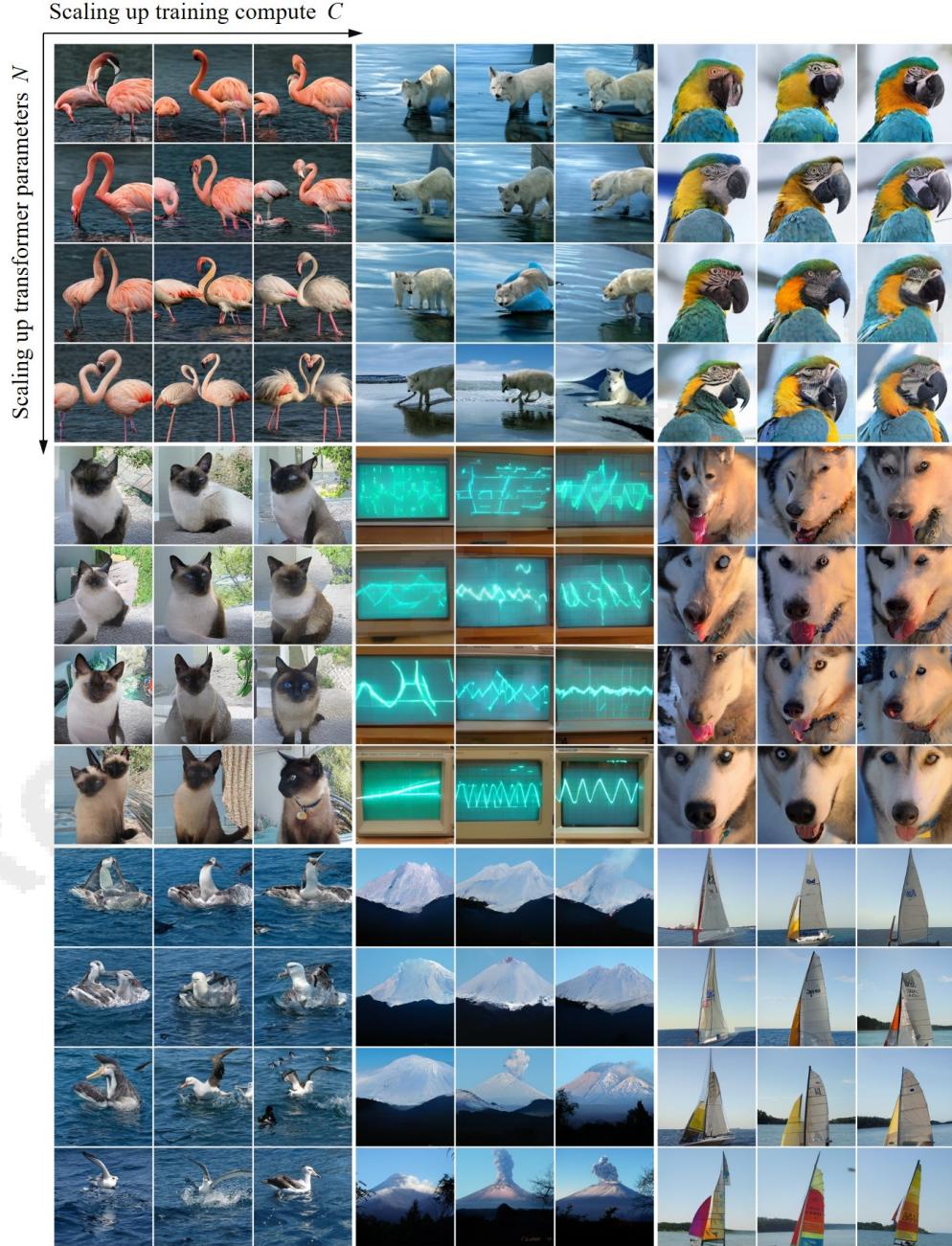


图 7：缩放模型大小  $N$  和训练计算  $C$  提高了视觉保真度和健全性。放大

为了更好的视野。样本是从 4 个不同规模和 3 个不同训练阶段的 VAR 模型中抽取的。9类标签（从左到右，从上到下）分别是：火烈鸟130、北极狼270、金刚鹦鹉88、暹罗猫284、示波器688、哈士奇250、mollymawk 146、火山980和双体船484。

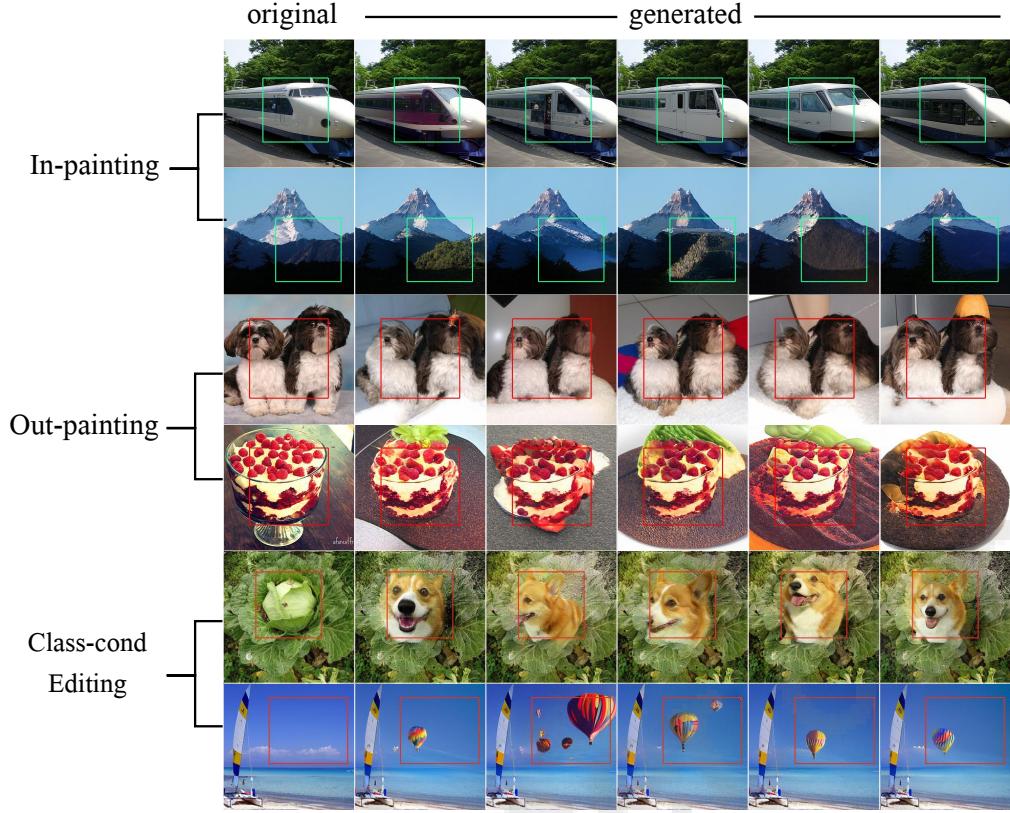


图 8：包含 in-painting、out-painting 和 class- 的下游任务中的零样本评估  
条件编辑。结果表明，VAR 可以推广到新的下游任务，无需特殊设计和微调。放大以获得更好的视图。

## B 零样本任务泛化

图像内画和外画。VAR-d30 经过测试。对于内画和外画，我们在蒙版外强制地面真值标记，并让模型仅生成蒙版内的标记。没有类标签信息被注入到模型中。结果如图 8 所示。在不修改网络架构或调整参数的情况下，VAR 在这些下游任务上取得了不错的结果，证实了 VAR 的泛化能力。

类条件图像编辑。继 MaskGIT [17] 之后，我们还在类条件图像编辑任务上测试了 VAR。与修复的情况类似，模型被迫仅在某些类标签条件下的边界框中生成令牌。图 8 显示该模型可以生成与周围环境很好融合的合理内容，再次验证了 VAR 的通用性。

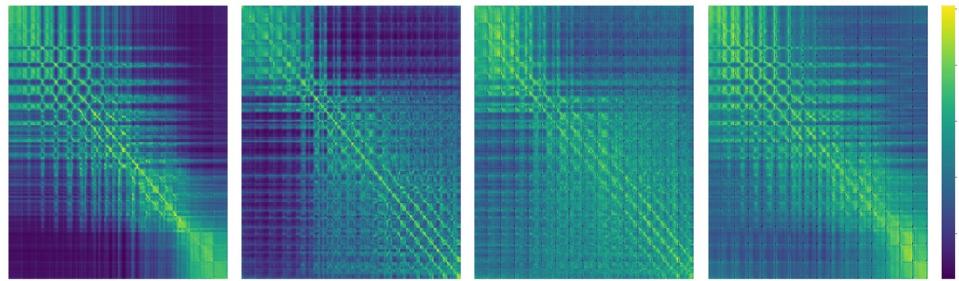


图 9：绘制了令牌依赖性。最后一个 self-attention 中注意力分数的归一化热图  
VQGAN 编码器层可视化。使用来自 ImageNet 验证集的 4 个随机  $256 \times 256$  图像。

### VQVAE 中的 C 令牌依赖项

为了检查 VQVAE [30] 中的标记依赖性，我们检查矢量量化模块之前的自注意力层中的注意力分数。我们从 ImageNet 验证集中随机采样  $4 \times 256 \times 256$  图像用于此分析。请注意，[30] 中的自注意力层只有 1 个头，因此对于每个图像，我们只绘制一个注意力图。图 9 中的热图显示了每个 token 相对于所有其他 token 的注意力分数，这表明所有 token 之间存在很强的双向依赖性。这并不奇怪，因为经过训练来重建图像的 VQVAE 模型利用自注意力层，而无需任何注意力掩模。一些工作[87]在视频VAE的自注意力层中使用了因果注意力，但我们没有发现任何图像VAE工作使用因果自注意力。

### D AR和VAR生成的时间复杂度

我们证明了AR和VAR生成的时间复杂度。引理D.1。对于标准的自注意力变压器，AR生成的时间复杂度为  $O(n^6)$ ，其中  $h = w = n$  并且  $h, w$  分别是VQ码图的高度和宽度。

证明。令牌总数为  $h \times w = n^2$ 。对于第  $i$  ( $1 \leq i \leq n^2$ ) 次自回归迭代，需要计算每个 token 与所有其他 token 之间的注意力分数，这需要  $O(i^2)$  时间。所以总时间复杂度为：

$$\sum_{i=1}^{n^2} i^2 = \frac{1}{6} n^2 (n^2 + 1) (2n^2 + 1), \quad (17)$$

这相当于  $O(n^6)$  基本计算。  $\square$

对于VAR，需要我们定义自回归生成的解析序列  $(h_1, w_1, h_2, w_2, \dots, h_K, w_K)$ ，其中  $h_i, w_i$  是第  $i$  个自回归步的 VQ 码图， $h_K = h, w_K = w$  达到最终分辨率。为简单起见，假设  $n_k = h_k = w_k$  对于所有  $1 \leq k \leq K$  且  $n = h = w$ 。我们将分辨率设置为  $n_k = a^{(k-1)}$ ，其中  $a > 1$  是一个常数，使得  $a^{(K-1)} = n$ 。引理 D.2。对于标准自注意力变压器和给定的超参数  $a > 1$ ，VAR 生成的时间复杂度为  $O(n^4)$ ，其中  $h = w = n$  且  $h, w$  是最后一个（最大）VQ 代码图的高度和宽度，分别。

证明。考虑第  $k$  ( $1 \leq k \leq K$ ) 个自回归生成。当前所有令牌映射  $(r_1, r_2, \dots, r_k)$  的令牌总数为：

$$\sum_{i=1}^k n_i^2 = \sum_{i=1}^k a^{2 \cdot (k-1)} = \frac{a^{2k} - 1}{a^2 - 1}. \quad (18)$$

因此，第  $k$  个自回归生成的时间复杂度为：

$$\left( \frac{a^{2k} - 1}{a^2 - 1} \right)^2. \quad (19)$$

通过总结所有自回归代，我们有：

$$\sum_{k=1}^{\log_a(n)+1} \left( \frac{a^{2k} - 1}{a^2 - 1} \right)^2 \quad (20)$$

$$= \frac{(a^4 - 1) \log n + (a^8 n^4 - 2a^6 n^2 - 2a^4 (n^2 - 1) + 2a^2 - 1) \log a}{(a^2 - 1)^3 (a^2 + 1) \log a} \quad (21)$$

$$\sim \mathcal{O}(n^4). \quad (22)$$

这样就完成了证明。  $\square$

BigGAN (FID=6.95)      VQVAE-2 (FID=31)      MaskGIT (FID=6.18)      VAR, ours (FID=1.92)

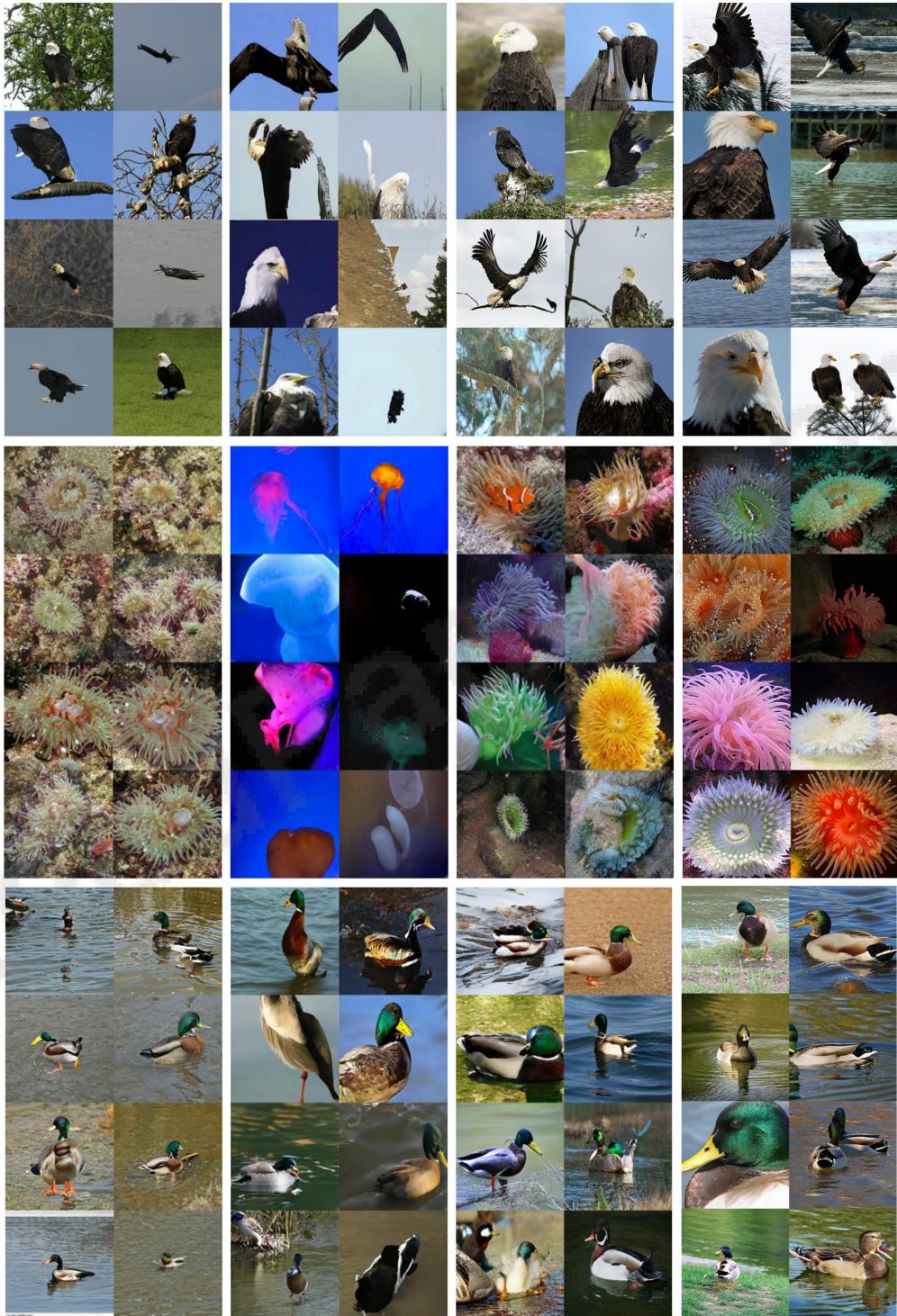


图 10: ImageNet 256×256 基准上的模型比较。更多生成的 512×512 样本  
VAR 可以在提交的补充材料 zip 文件中找到。

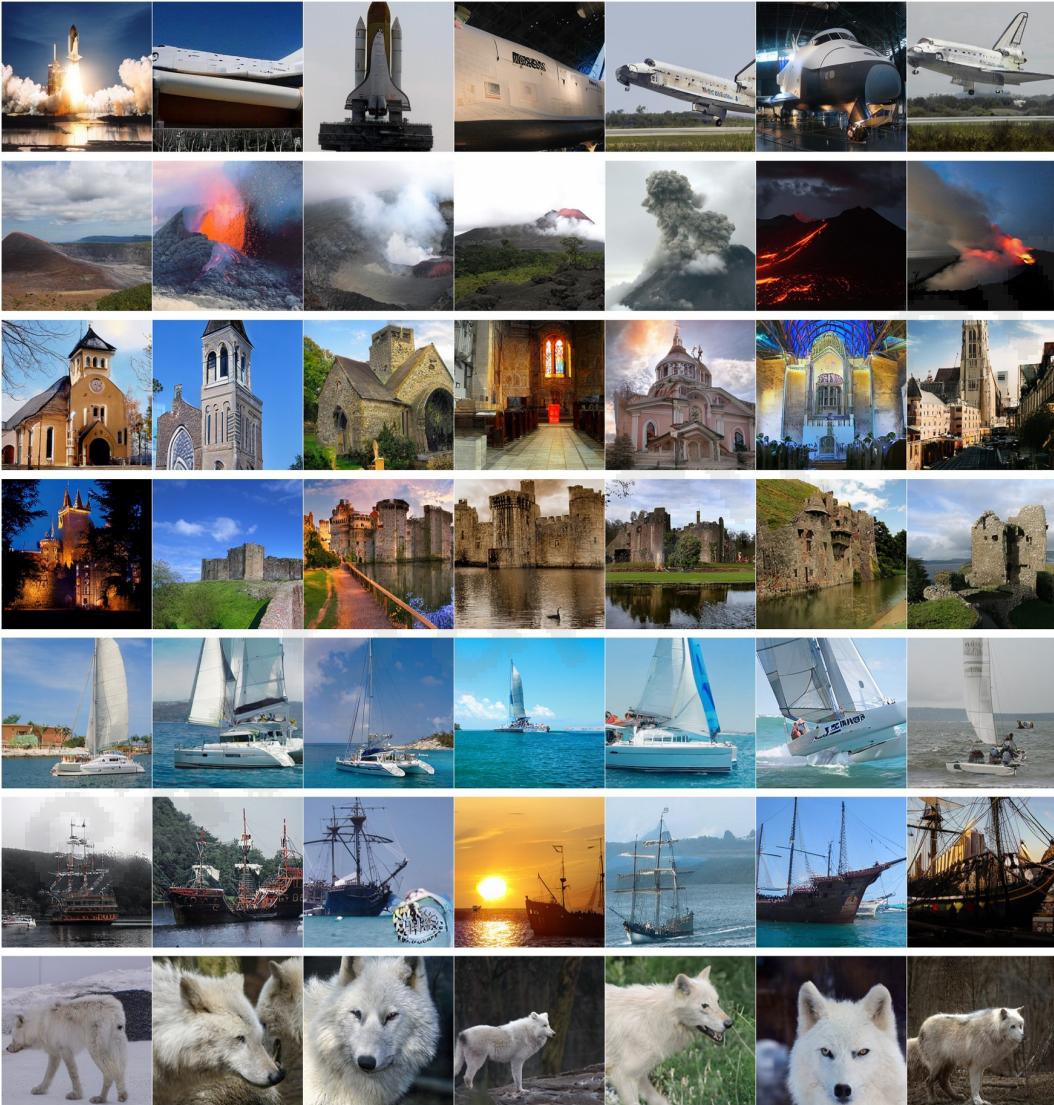


图 11：一些通过在 ImageNet 上训练的 VAR 生成  $256 \times 256$  样本。更多生成  $512 \times 512$  VAR 提供的样本可以在提交的补充材料 zip 文件中找到。

## References

- [1] J. Achiam、S. Adler、S. Agarwal、L. Ahmad、I. Akkaya、F. L. Aleman、D. Almeida、J. Altenschmidt、S. 奥特曼, S. 阿纳德卡特, 等人。Gpt-4 技术报告。arXiv 预印本 arXiv:2303.08774, 2023. 2, 3, 8, 9[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. 雷诺兹等人。Flamingo: 用于小样本学习的视觉语言模型。神经信息处理系统的进展, 35: 23716–23736, 2022. 3[3] Alpha-VLLM。大点图像网。https://github.com/Alpha-VLLM/LLAMA2-Accessory/tree/f7fe19834b23e38f333403b91bb0330afe19f79e/Large-DiT-ImageNet, 2024. 2, 7, 8[4] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. 莱皮欣, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen 等人。Palm 2 技术报告。arXiv 预印本 arXiv:2305.10403, 2023. 2[5] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. 黄, 等人。Qwen 技术报告。arXiv 预印本 arXiv:2309.16609, 2023. 2[6] Y. Bai, X. Geng, K. Mangalam, A. Bar, A. Yuille, T. Darrell, J. Malik 和 A. A. Efros。顺序建模支持大型视觉模型的可扩展学习。arXiv 预印本 arXiv:2312.00785, 2023. 2, 3[7] F. Bao, C. Li, J. Zhu, 和 B. Zhang。Analytic-dpm: 最佳反向方差扩散概率模型的分析估计。arXiv 预印本 arXiv:2201.06503, 2022. 3[8] F. 包, S. 聂, K. 薛, Y. 曹, C. 李, H. 苏, J. 朱。所有这些都值得一提: 扩散模型的 vit 支柱。IEEE/CVF 计算机视觉与模式识别会议论文集, 第 22669–22679 页, 2023. 3[9] F. Bao, C. Shang, G. Yue, G. He, H. Zhu, K. Cheng, M. 赵, S. 刘, Y. 王, 和 J. 朱。Vidu: 高度一致、动态且熟练的文本到视频生成器, 具有扩散模型。arXiv preprint arXiv:2405.04233, 2024. 3[10] H. Bao, L. Dong, S. Piao, and F. Wei。Beit: 图像转换器的 Bert 预训练。arXiv preprint arXiv:2106.08254, 2021. 3[11] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson 和 A. Efros。通过图像修复进行视觉提示。神经信息处理系统的进展, 35:25005–25017, 2022. 3[12] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli, et al. Lumiere: 视频时空扩散模型一代。arXiv 预印本 arXiv:2401.12945, 2024. 3[13] A. Brock, J. Donahue 和 K. Simonyan。用于高保真自然图像合成的大规模 gan 训练。arXiv 预印本 arXiv:1809.11096, 2018. 7[14] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. 泰勒, T. 卢曼, E. 卢曼, C. Ng, R. Wang 和 A. Ramesh。视频生成模型作为世界模拟器。OpenAI, 2024. 3, 7, 10[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. 阿斯克尔等人。语言模型是小样本学习者。神经信息处理系统的进展, 33:1877–1901, 2020. 2, 3[16] H. Chang, H. Zhu, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. 杨, K. 墨菲, W. T. 弗里曼, M. 鲁宾斯坦等人。Muse: 通过屏蔽生成转换器生成文本到图像。arXiv preprint arXiv:2301.00704, 2023. 3[17] H. Chang, H. 张, L. Jiang, C. Liu 和 W. T. Freeman。Maskgit: 掩模生成图像转换器。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 11315–11325 页, 2022. 3, 7, 9, 12[18] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu 和 Z. Li。Pixart-\sigma: 用于 4k 文本到图像生成的扩散变换器的弱到强训练。arXiv 预印本 arXiv:2403.04692, 2024. 3[19] 陈建, 于建, 葛成, 姚丽, 谢鄂, 吴勇, 王志, 郭建, 罗鹏, 陆红, 等。Pixart: 扩散变压器的快速训练, 用于逼真的文本到图像合成。arXiv 预印本 arXiv:2310.00426, 2023. 3, 6[20] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan 和 I. Sutskever。来自像素的生成预训练。国际机器学习会议, 第 1691–1703 页。PMLR, 2020. 3[21] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, Z. Muyan, Q. Zhu, X. Zhu, L. Lu, et al。实习: 扩大视觉基础模型并调整通用视觉语言任务。arXiv preprint arXiv:2312.14238, 2023. 3[22] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. 格尔曼等人。Palm: 通过路径扩展语言建模。机器学习研究杂志, 24(240):1–113, 2023. 2[23] X. Dai, J. Hou, C.-Y. 马, S. 蔡, J. 王, R. 王, P. 张, S. Vandenende, X. 王, A. Dubey, 等。Emu: 利用大海捞针增强图像生成模型。arXiv preprint arXiv:2309.15807, 2023. 3[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li 和 L. Fei-Fei。Imagenet: 大规模分层图像数据库。2009 年 IEEE 计算机视觉和模式识别会议, 第 248–255 页。IEEE, 2009. 8, 21, 22

- [25] J.德夫林, M.-W。张、K. Lee 和 K. Toutanova。Bert: 用于语言理解的深度双向变压器的预训练。arXiv 预印本 arXiv:1810.04805, 2018. 3[26] P. Dhariwal 和 A. Nichol。扩散模型在图像合成方面击败了甘斯。神经信息处理系统进展, 34:8780–8794, 2021. 7[27] R. Dong, C. Han, Y. Peng, Z. Qi, Z. Ge, J. Yang, L. Zhao, J. Sun, H.周, H.魏, 等人。Dreamilm: 协同多模态理解和创造。arXiv 预印本 arXiv:2309.11499, 2023. 3[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Min-derer, G. Heigold, S.Gelly 等人。一张图像相当于 16x16 个单词: 用于大规模图像识别的 Transformers。arXiv 预印本 arXiv:2010.11929, 2020. 3[29] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F.博塞尔, D.波德尔、T. Dockhorn, Z. English、K. Lacey、A. Goodwin, Y. Marek 和 R. Rombach。用于高分辨率图像合成的缩放整流流变压器, 2024. 3, 7[30] P. Esser, R. Rombach 和 B. Ommer。驯服变压器以进行高分辨率图像合成。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 12873-12883 页, 2021 年。2, 3, 4, 5, 6, 7, 10, 13[31] 高圣, 周平, M.-M.程和S.严。Mdvt2: Masked Diffusion Transformer 是一个强大的图像合成器。arXiv 预印本 arXiv:2303.14389, 2023. 3[32] Y. Ge, S. Zhao, Z. Zeng, Y. Ge, C. Li, X. Wang, and Y. Shan。使用seedtokenizer 让美洲驼看到并画画。arXiv 预印本 arXiv:2310.01218, 2023. 3[33] Y. Ge, S. Zhao, J. Zhu, Y. Ge, K. Yi, L. Song, C. Li, X. Ding, and Y. Shan。Seed-x: 具有统一多粒度理解和生成的多模态模型。arXiv 预印本 arXiv:2404.14396, 2024. 3[34] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei-Fei, I. Essa, L. Jiang, and J. Lezama。使用扩散模型生成逼真的视频。arXiv 预印本 arXiv:2312.06662, 2023. 3[35] K. He、X. Chen、S. Xie、Y. Li、P. Dollár 和 R. Girshick。屏蔽自动编码器是可扩展的视觉学习器。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 16000-16009 页, 2022 年。3[36] T. Henighan, J. Kaplan, M. Katz、M. Chen、C. Hesse、J. Jackson、H Jun、T. B. Brown、P. Dhariwal、S. Gray 等人。自回归生成模型的缩放定律。arXiv 预印本 arXiv:2010.14701, 2020. 2, 3, 8, 9[37] J. Ho、C. Saharia、W. Chan、D. J. Fleet、M. Norouzi 和 T. Salimans。用于生成高保真图像的级联扩散模型。机器学习研究杂志, 23(1):2249–2281, 2022. 3, 7[38] J. Ho 和 T. Salimans。无分类器的扩散指导。arXiv 预印本 arXiv:2207.12598, 2022. 3[39] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. 卡萨斯、L.A.亨德里克斯、J.韦尔布尔、A.克拉克等人。训练计算优化的大型语言模型。arXiv preprint arXiv:2203.15556, 2022. 2, 3, 8, 9[40] L. Huang, W. Yu, W. Ma, W.zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X冯, 秦, 等。大语言模型中的幻觉调查: 原则、分类、挑战和开放问题。arXiv preprint arXiv:2311.05232, 2023. 2[41] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie、B. Hariharan 和 S.-N. 林.视觉提示调整。欧洲计算机视觉会议, 第 709-727 页。Springer, 2022. 3[42] Y. Jin, K. Xu, L. Chen, C. Liao, J. Tan, B. Chen, C. Lei, A. Liu, C. Song, X. Lei, et al。具有动态离散视觉标记化的统一语言视觉预训练。arXiv 预印本 arXiv:2309.04669, 2023.3[43] M. Kang, J.-Y. 朱、R. 张、J. Park、E. Shechtman、S. Paris 和 T. Park。扩展甘斯以进行文本到图像的合成。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 10124–10134 页, 2023. 6, 7[44] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess、R. Child、S. Gray、A. Radford、J. Wu 和 D.阿莫代伊。神经语言模型的缩放定律。arXiv 预印本 arXiv:2001.08361, 2020.2,3,6,8,9[45] T. Karras, T. Aila、S. Laine 和 J. Lehtinen。甘斯的渐进生长可提高质量、稳定性和变化。arXiv 预印本 arXiv:1710.10196, 2017. 2[46] T. Karras, M. Aittala、S. Laine、E. Härkönen、J. Hellsten、J. Lehtinen 和 T. Aila。无别名生成对抗网络。神经信息处理系统进展, 34:852–863, 2021. 6[47] T. Karras, S. Laine 和 T. Aila。用于生成对抗网络的基于样式的生成器架构。《IEEE/CVF 计算机视觉和模式识别会议论文集》, 第 4401-4410 页, 2019 年。4, 6[48] T. Karras, S. Laine、M. Aittala、J. Hellsten、J. Lehtinen 和 T. Aila。stylegan的图像质量分析与改进。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 8110–8119 页, 2020 年。6[49] A. Kirillov、E. Mintun、N. Ravi、H. Mao、C. Rolland、L. Gustafson、T肖, S.怀特海德, A.C.伯格, W.-Y. 罗, 等人。分割任何东西。arXiv 预印本 arXiv:2304.02643, 2023. 3[50] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallochi, 一个。科列斯尼科夫等人。开放图像数据集 v4: 大规模统一图像分类、对象检测和视觉关系检测。国际计算机视觉杂志, 128 (7) : 1956–1981, 2020. 6

- [51] D. Lee、C. Kim、S. Kim、M. Cho 和 W.-S. 韩。使用残差量化的自回归图像生成。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 11523–11532 页, 2022 年。
- 2,3,4,5,6,7[52] T. Li、D. Katabi 和 K. He。通过生成表示来生成自条件图像。arXiv preprint arXiv:2312.03701, 2023. 2, 7[53] T.-Y. Lin、P. Dollár、R. Girshick、K. He、B. Hariharan 和 S. Belongie。用于目标检测的特征金字塔网络。IEEE 计算机视觉和模式识别会议论文集, 第 2117–2125 页, 2017 年。2[54] H. Liu、C. Li、Q. Wu 和 Y. J. Lee。视觉指令调整。神经信息处理系统的进展, 36, 2024。3[55] D. G. Lowe。从局部尺度不变特征进行对象识别。第七届 IEEE 国际计算机视觉会议论文集, 第 2 卷, 第 1150–1157 页。Ieee, 1999。2[56] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu。Dpm-solver: 一种快速求解器, 用于扩散概率模型采样, 大约需要 10 个步骤。神经信息处理系统的进展, 35: 5775–5787, 2022。3[57] 陆成, 周勇, 包芳, 陈建, 李成, 朱建。Dpm-solver++: 用于扩散概率模型引导采样的快速求解器。arXiv 预印本 arXiv:2211.01095, 2022. 3[58] J. Lu、C. Clark、S. Lee、Z. Zhang、S. Khosla、R. Marten、D. Hoiem 和 A. Kembhavi。Unified-io2: 通过视觉、语言、音频和动作扩展自回归多模态模型。arXiv preprint arXiv:2312.17172, 2023. 2[59] J. Lu、C. Clark、R. Zellers、R. Mottaghi 和 A. Kembhavi。Unified-io: 视觉、语言和多模态任务的统一模型。arXiv 预印本 arXiv:2206.08916, 2022. 2[60] F. Mentzer、D. Minnen、E. Agustsson 和 M. Tschannen。有限标量量化: Vq-vae 制作简单。arXiv 预印本 arXiv:2309.15505, 2023. 10[61] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, 和陈先生。Glide: 使用文本引导扩散模型生成和编辑逼真的图像。arXiv preprint arXiv:2112.10741, 2021. 3[62] M. Oquab, T. Dariseti, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. 马萨, A. El-Nouby, 等人。Dinov2: 在没有监督的情况下学习强大的视觉特征。arXiv preprint arXiv:2304.07193, 2023. 3[63] L. 欧阳, J. 吴, X. 江, D. Almeida, C. Wainwright, P. Mishkin, C. 张, S. Agarwal, K. Slama, A. 雷等人。训练语言模型遵循人类反馈的指令。神经信息处理系统进展, 35: 27730–27744, 2022。2[64] W. Peebles 和 S. Xie。具有变压器的可扩展扩散模型。IEEE/CVF 国际计算机视觉会议论文集, 第 4195–4205 页, 2023 年。2,3,6,7,8[65] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. 阿加瓦尔, G. 萨斯特里, A. 阿斯克尔, P. 米什金, J. 克拉克等人。从自然语言监督中学习可迁移的视觉模型。国际机器学习会议, 第 8748–8763 页。PMLR, 2021. 3[66] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever 等人。通过生成预训练提高语言理解。文章, 2018。2[67] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever 等。语言模型是无监督的多任务学习者。OpenAI 博客, 1(8):9, 2019.
- 2,3,6[68] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen 和我苏茨克韦尔。零镜头文本到图像生成。国际机器学习会议, 第 8821–8831 页。PMLR, 2021。2[69] A. Razavi, A. Van den Oord 和 O. Vinyals。使用 vqvae-2 生成多样化的高保真图像。神经信息处理系统的进展, 32, 2019. 2,3,7[70] S. Reed, A. Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov 和 N. Freitas。并行多尺度自回归密度估计。国际机器学习会议, 第 2912–2921 页。PMLR, 2017. 3[71] R. Rombach, A. Blattmann, D. Lorenz, P. Esser 和 B. Ommer。使用潜在扩散模型进行高分辨率图像合成。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 10684–10695 页, 2022. 3, 7[72] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan、T. Salimans 等人。具有深度语言理解的逼真文本到图像扩散模型。神经信息处理系统进展, 35: 36479–36494, 2022. 3[73] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja 等人。多任务提示训练可实现零样本任务泛化。arXiv preprint arXiv:2110.08207, 2021. 3[74] A. Sauer, T. Karras, S. Laine, A. Geiger 和 T. Aila。Stylegan-t: 释放 gan 的力量, 实现快速大规模文本到图像的合成。arXiv 预印本 arXiv:2301.09515, 2023. 6[75] A. Sauer, K. Schwarz 和 A. Geiger。Stylegan-xl: 将 stylegan 扩展到大型多样化数据集。ACMSIGGRAPH 2022 会议记录, 第 1–10 页, 2022. 6, 7

- [76] J.宋, C.孟, 和S.埃尔蒙。去噪扩散隐式模型。arXiv 预印本 arXiv: 2010.02502, 2020. 3[77] Y.宋和S.埃尔蒙。通过估计数据分布的梯度进行生成建模。神经信息处理系统进展, 32, 2019. 3[78] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gau, J. Liu, T. Huang, 和X.王。多模态生成预训练。arXiv 预印本 arXiv:2307.05222, 2023. 3[79] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y.卢, 等。Ernie 3.0: 大规模知识增强语言理解和生成的预训练。arXiv preprint arXiv:2107.02137, 2021. 2[80] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth 等人。Gemini: 一系列高性能的多式联运模型。arXiv 预印本 arXiv:2312.11805, 2023. 2[81] C. Tian, X. Zhu, Y. Xiong, W. Wang, Z. Chen, W. Wang, Y. Chen, L. Lu, T. Lu, J.周, 等人。Mm-interleaved: 通过多模态特征同步器进行交错图像文本生成建模。arXiv preprint arXiv:2401.10208, 2024. 3[82] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, 和Z. Yuan。为卷积网络设计 bert: 稀疏和分层掩码建模。arXiv 预印本 arXiv:2301.03580, 2023. 2[83] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F.阿扎尔等人。Llama: 开放高效的基础语言模型。arXiv 预印本 arXiv: 2302.13971, 2023. 2, 3, 6[84] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaie, N. Bashlykov, S. Batra, P. Bhargava, S.博萨莱等人。Llama 2: 开放基础和微调的聊天模型。arXiv 预印本 arXiv: 2307.09288, 2023. 2, 3, 6[85] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves 等。使用 PixelCNN 解码器生成条件图像。神经信息处理系统的进展, 29, 2016. 3[86] A. Van Den Oord, O. Vinyals 等。神经离散表示学习。神经信息处理系统的进展, 30, 2017. 3[87] R. Villegas, M. Babaeizadeh, P.-J. 金德曼斯, H. Moraldo, H. 张、M. T. Saffar, S. Castro, J. Kunze 和 D. Erhan。Phenaki: 根据开放域文本描述生成可变长度视频。国际学习表征会议, 2022. 13[88] H. Wang, H. Tang, L. Jiang, S. Shi, M. F. Naeem, H. Li, B. Schiele, and L. Wang。Git: 通过通用语言界面实现通才愿景转换器。arXiv 预印本 arXiv:2403.09394, 2024. 3[89] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. 乔, 等人。Visionllm: 大型语言模型也是一个用于以视觉为中心的任务的开放式解码器。神经信息处理系统进展, 36, 2024. 3[90] X. Wang, W. Wang, Y. Cao, C. Shen, 和 T. Huang。图像以图像说话: 一位进行情境视觉学习的通才画家。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 6830–6839 页, 2023 年。3[91] B. Workshop, T. L. Scao、A. Fan, C. Akiki, E. Pavlick, S. Ilić, D.赫斯洛, R.卡斯塔尼, A.S.卢乔尼, F.伊冯等。Bloom: 176b 参数的开放获取多语言语言模型。arXiv preprint arXiv:2211.05100, 2022. 2, 3[92] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. qin, A. Ku, Y. Xu, J. Baldridge, and Y.吴。使用改进的 vqgan 进行矢量量化图像建模。arXiv 预印本 arXiv:2110.04627, 2021. 2, 3, 4, 6, 7[93] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. 内容丰富的文本到图像的缩放自回归模型一代。arXiv 预印本 arXiv:2206.10789, 2(3):5, 2022. 3[94] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. 杨, Y. 浩, I. Essa, 等。Magvit: 蒙面生成视频转换器。IEEE/CVF 计算机视觉和模式识别会议论文集, 第 10459–10469 页, 2023. 3[95] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann 等人。语言模型胜过扩散——分词器是视觉生成的关键。arXiv preprint arXiv:2310.05737, 2023. 3, 10[96] L. Yu, B. Shi, R. Pasunuru, B. Muller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S.谢宁等人。缩放自回归多模态模型: 预训练和指令调整。arXiv preprint arXiv:2309.02591, 2(3), 2023. 3[97] R. 张、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang。深度特征作为感知指标的不合理有效性。IEEE 计算机视觉和模式识别会议论文集, 第 586–595 页, 2018 年。4[98] S. 张、S. Roller、N. Goyal、M. Artetxe、M. Chen、S. Chen、C. Dewan、M. Diab、X. Li、X. V. Lin 等人。Opt: 开放预训练的 Transformer 语言模型。arXiv 预印本 arXiv:2205.01068, 2022. 3[99] C. Cheng, T.-L. Vuong, J. Cai 和 D. Phung。Movq: 调制量化矢量以生成高保真图像。神经信息处理系统的进展, 35:23412–23425, 2022. 2, 10

1. 索赔

问题：摘要和引言中提出的主要主张是否准确反映了论文的贡献和范围？

Answer: [Yes]

理由：是的。我们的主要贡献也在第 2 节中有详细介绍。1. 另请参见第 1 节。5 和附录 D 了解更多理论和实验证据。

Guidelines:

- 答案NA 表示摘要和引言不包括论文中提出的主张。
- 摘要和/或引言应清楚地说明所提出的主张，包括论文中所做的贡献以及重要的假设和限制。对于这个问题，“否”或“NA”的答案不会被审稿人很好地理解。
- 所提出的主张应与理论和实验结果相匹配，并反映结果在多大程度上可以推广到其他环境。
- 可以将理想目标包括为只要论文明确没有实现这些目标，就没有动机。

2. 局限性

问题：本文是否讨论了作者所做工作的局限性？

Answer: [Yes]

理由：是的，请参阅第 2 节。7 的限制。我们还报告了很多有关计算效率的内容，例如表中所示。1 和附录 D。

Guidelines:

- 答案NA 表示论文没有限制，而答案No 表示论文有局限性，但论文中未讨论这些局限性。
- 鼓励作者在论文中创建单独的“局限性”部分。
- 论文应指出任何强有力假设以及结果对于违反这些假设的鲁棒性如何（例如，独立假设、无噪声设置、模型良好规范、仅在局部保持的渐近近似）。作者应该反思这些假设在实践中如何被违反以及会产生什么影响。作者应该反思所提出的主张的范围，例如，如果该方法仅在几个数据集或几次运行中进行了测试。一般来说，经验结果通常依赖于隐含的假设，这些假设应该被阐明。作者应该反思影响该方法性能的因素。例如，当图像分辨率低或图像拍摄时，面部识别算法可能会表现不佳。低照明。或者，语音转文本系统可能无法可靠地为在线讲座提供隐藏式字幕，因为它无法处理技术术语。
- 作者应讨论所提出算法的计算效率以及它们如何随数据集大小进行扩展。
- 如果适用，作者应该讨论他们解决隐私和公平问题的方法可能存在的局限性。
- 虽然作者可能担心审稿人可能会利用完全诚实的限制作为拒绝的理由，但更糟糕的结果可能是审稿人发现审稿人未承认的局限性。
- 纸。作者应该运用他们的最佳判断，并认识到有利于透明度的个人行动在制定维护社区完整性的规范方面发挥着重要作用。将特别指示审稿人不要惩罚有关限制的诚实行为。

3 理论假设与证明

问题：对于每个理论结果，论文是否提供了全套假设和完整（且正确）的证明？

答案：[是]理由：我们在附录D中详细介绍了时间复杂度的理论结果的假设和证明。指南：· 答案NA表示论文不包含理论结果。· 论文中的所有定理、公式和证明应该编号和交叉引用。· 所有假设都应该在任何定理的陈述中清楚地陈述或引用。· 证明可以出现在主论文或补充材料中，但如果它们出现在补充材料中 ss-  
材料，鼓励作者提供简短的草图以提供直觉。· 相反，论文核心提供的任何非正式证明应辅以附录或补充材料中提供的正式证 s.  
明。· 证明所依赖的定理和引理应该正确引用。4.实验结果再现 if  
性问题：论文是否充分披露了再现论文主要实验结果所需的所有 ort  
信息，其程度影响了论文的主要主张和/或结论（无论是否提供了 ed  
代码和数据或不是）？答案：[是]理由：我们使用可公开访问的  
数据集 ImageNet [24]。我们上传代码和说明以恢复结果。指  
南：· 答案 NA 表示论文不包含实验。· 如果论文包含实验， ex-  
则审稿人不会很好地理解此问题的“否”答案：使论文具有可重 ns  
复性无论是否提供代码和数据，这一点都很重要。· 如果贡献是 nd  
数据集和/或模型，作者应描述为使其结果可重现或可验证而采取  
的步骤。· 根据贡献，可重复性可以通过多种方式实现。例如， ed  
如果贡献是一种新颖的架构，充分描述该架构就足够了，或者如 of  
果贡献是特定的模型和经验评估，则可能有必要使其他人能够复 en  
制该架构具有相同数据集的模型，或提供对模型的访问。一般来说  
ys.  
发布代码和数据通常是实现此目的的一个好方法，但是也可以  
lly  
通过如何复制结果、访问托管模型（例如，在大型语言模型的  
ay  
情况下）、发布模型检查点的详细说明来提供可重复性，或其他  
e  
适合所进行研究的方式。· 虽然 NeurIPS 不要求发布代码，但  
en  
会议确实要求所有提交内容提供一些合理的再现性途径，这可能  
ed  
取决于贡献的性质。例如 (a) 如果贡献主要是一个新算法，论  
se  
文应该明确如何重现该算法。 (b) 如果贡献主要是一个新模型  
re  
架构，论文应该清楚、完整地描述该架构。 (c) 如果贡献的是  
is-  
一个新模型（例如，一个大型语言模型），那么应该有一种方法  
he  
来访问该模型以重现结果，或者一种重现模型的方法（例如，使  
ow  
用开源数据集或有关如何构建(d) 我们认识到，在某些情况下，  
be  
再现性可能很棘手，在这种情况下，欢迎作者描述他们提供再现  
ld  
性的特定方式。在闭源模型的情况下，对模型的访问可能是以某  
ce  
种方式受到限制（例如，注册用户），但其他研究人员应该有可  
ct  
能有某种途径来复制或验证结果。5. 开放数据和代码的访问

问题：论文是否提供对数据和代码的开放访问，并提供足够的说明来忠实地再现主要实验结果，如补充材料中所述？

Answer: [Yes]

理由：我们使用可公开访问的数据集 ImageNet [24]。我们上传代码和说明以恢复结果。盲审期结束后，我们将开源所有代码、指令和模型检查点。

Guidelines:

- 答案 NA 表示论文不包含需要代码的实验。 · 请参阅 NeurIPS 代码和数据提交指南 (<https://nips.cc/public/guides/CodeSubmissionPolicy>) 了解更多详细信息。 · 虽然我们鼓励发布代码和数据，我们知道这可能是不可能的，所以“否”是一个可以接受的答案。论文不能仅仅因为不包含代码而被拒绝，除非这是贡献的核心（例如，对于新的开源基准）。 · 说明应包含运行以重现结果所需的确切命令和环境。有关更多详细信息，请参阅 NeurIPS 代码和数据提交指南 (<https://nips.cc/public/guides/CodeSubmissionPolicy>)。 · 作者应提供有关数据访问和准备的说明，包括如何访问原始数据、预处理数据、中间数据数据和生成的数据等。 · 作者应提供脚本来重现新提出的方法和基线的所有实验结果。如果只有一部分实验是可重复的，他们应该说明脚本中省略了哪些实验以及原因。 · 在提交时，为了保持匿名，作者应发布匿名版本（如果适用）。 · 在补充材料中提供尽可能多的信息建议（附加到本文中），但允许包含数据和代码的 URL。

## 6. Experimental Setting/Details

问题：论文是否指定了理解结果所需的所有训练和测试细节（例如，数据分割、超参数、如何选择它们、优化器的类型等）？

Answer: [Yes]

理由：请参阅第 2 节。5 和附录 4。

Guidelines:

- 答案 NA 表示论文不包含实验。 · 实验设置应在论文的核心部分详细介绍，以了解结果并理解结果。 · 可以提供完整的详细信息可以在附录中包含代码，也可以作为补充材料。

## 7. 实验统计意义

问题：论文是否报告了适当且正确定义的误差线或有关实验统计显着性的其他适当信息？

Answer: [No]

理由：由于资源限制，我们不报告误差线。请注意 inSec。5 我们为扩展规律研究花费了大量资源（我们训练了 12 个不同的模型），这使得多次运行每个实验变得令人望而却步。

Guidelines:

- 答案 NA 表示论文不包含实验。 · 如果结果附有误差线、置信区间或统计显着性检验，至少对于支持主要主张的实验，作者应回答“是”纸。

- 应明确说明误差线捕获的变异因素（例如，训练/测试分割、初始化、随机绘制某些参数或给定实验条件下的总体运行）。 · 应解释计算误差线的方法（封闭式公式、调用库函数、引导程序等） · 应给出所做的假设（例如，正态分布误差）。 · 应明确误差线是标准差还是平均值的标准误差。 · 可以举报1-sigma 误差线，但应该说明这一点。如果误差正态性假设未得到验证，作者最好报告 2-sigma 误差条，而不是声明其具有 96% CI。 · 对于非对称分布，作者应注意不要在表格或图中显示对称误差条：会产生超出范围的结果（例如负错误率）。 · 如果在表格或图中报告了误差线，作者应在文本中解释它们的计算方法，并参考图中相应的数字或表格。
- 文本.8. 实验计算资源问题：对于每个实验，论文是否提供了重现实验所需的计算机资源（计算工作者类型、内存、执行时间）的足够信息？答案：[是]理由：我们报告了训练 PFlops 图 6 和表中的速度。1 和选项卡。
2. 指导原则：
- 答案 NA 表示论文不包含实验。
  - 论文应指出计算工作者 CPU 或 GPU、内部集群或云提供商的类型，包括相关内存和存储。
  - 论文应提供每个实验运行所需的计算量以及估计总计算量。论文应披露整个研究项目是否需要比论文中报告的实验更多的计算量（例如，未进入研究阶段的初步或失败的实验）
- .9. 道德准则问题：本文中进行的研究在各个方面都符合 NeurIPS 道德准则 <https://neurips.cc/public/EthicsGuidelines>？答案：[是]理由：我们遵循 NeurIPS 道德准则。准则：
- 答案 NA 表示作者尚未审查 NeurIPS 道德准则。
  - 如果作者回答否，他们应解释需要偏离道德准则的特殊情况。
  - 作者应确保保持匿名（例如，如果由于其管辖范围内的法律或法规而有特殊考虑）。
10. 更广泛的影响问题：本文是否讨论了所开展工作的潜在积极社会影响和负面社会影响？答案：[否]理由：这项工作侧重于学术性、公开可用的基准 ImageNet。这项工作与任何私人或个人无关数据，并且不存在明显的负面影响。指南：
- 答案 NA 表示所执行的工作不存在社会影响。

· 如果作者回答“不适用”或“否”，他们应该解释为什么他们的工作没有社会影响，或者为什么论文没有解决社会影响。 · 负面社会影响的例子包括潜在的恶意或无意用途（例如，虚假信息、生成虚假个人资料、监视））、公平性考虑（例如，部署可能做出不公平地影响特定群体的决策的技术）、隐私考虑和安全考虑。 · 会议预计许多论文将是基础研究，不与特定应用程序绑定，更不用说部署了。然而，如果存在任何负面应用的直接路径，作者应该指出。例如，指出生模型质量的改进可用于生成虚假信息的深度伪造品是合理的。另一方面，无需指出优化神经网络的通用算法可以使人们能够更快地训练生成 Deepfakes 的模型。 · 作者应该考虑当该技术按预期使用并正常运行时可能出现的危害，当按预期使用该技术但给出不正确的结果时可能出现的危害，以及（有意或无意）滥用该技术而造成的危害。 · 如果存在负面影响，作者还可以讨论可能的缓解策略（例如，门控释放模型，除了攻击之外还提供防御，监控滥用的机制，监控系统如何从反馈中学习的机制，提高机器学习的效率和可访问性）。

#### 11. 保障措施

问题：本文是否描述了为负责任地发布滥用风险较高的数据或模型（例如预训练的语言模型、图像生成器或抓取的数据集）而采取的保障措施？

Answer: [No]

理由：我们预计这项工作不会有任何滥用的高风险。

Guidelines:

· 答案 NA 意味着该论文不存在此类风险。 · 已发布的具有高误用或双重用途风险的模型应在发布时采取必要的保障措施，以允许模型的受控使用，例如要求用户遵守使用指南或限制访问模型或实施安全过滤器。 · 从互联网上抓取的数据集可能会带来安全风险。作者应描述他们如何避免发布不安全图像。 · 我们认识到提供有效的保障措施具有挑战性，许多论文并不要求这样做，但我们鼓励作者考虑到这一点并尽最大努力。

#### 12. 现有资产的许可证

问题：论文中使用的资产（例如代码、数据、模型）的创建者或原始所有者是否得到了适当的认可，并且明确提及的许可和使用条款是否得到了适当的尊重？

Answer: [Yes]

理由：是的，我们以适当的方式归功于他们。

Guidelines:

· 答案NA 表示论文不使用现有资产。 · 作者应引用生成代码包或数据集的原始论文。 · 作者应说明使用了哪个版本的资产，如果可能，请包含URL。 · 每项资产均应包含许可证名称（例如 CC-BY 4.0）。 · 对于从特定来源（例如网站）抓取的数据，应提供该来源的版权和服务条款。

- 如果发布资产，则应提供包中的许可证、版权信息和使用条款。对于流行的数据集，[paperswithcode.com/datasets](http://paperswithcode.com/datasets) 为某些数据集提供了精选的许可证。他们的许可指南可以帮助确定数据集的许可。
- 对于重新打包的现有数据集，应提供原始许可和派生资产的许可（如果已更改）。
- 如果无法在线获取此信息，我们鼓励作者联系该资产的创建者。

#### 13. 新资产

问题：本文中引入的新资产是否有详细记录，并且文档是否与资产一起提供？

Answer: [NA]

理由：本文并未发布新资产。

Guidelines:

- 答案NA 意味着论文不会发布新资产。
- 研究人员应通过结构化模板将数据集/代码/模型的详细信息作为其提交内容的一部分进行交流。这包括有关培训、许可、限制等的详细信息。
- 论文应讨论是否以及如何获得资产使用者的同意。
- 提交时，请记住对您的资产进行匿名化（如果适用）。您可以创建匿名 URL 或包含匿名 zip 文件。

#### 14. 众包和人类受试者研究

问题：对于以人类为对象的众包实验和研究，论文是否包括向参与者提供的说明全文和屏幕截图（如果适用）以及有关补偿的详细信息（如果有）？

Answer: [NA]

理由：本文不涉及众包或人类受试者研究。

Guidelines:

- 答案 NA 表示论文不涉及众包，也不涉及人类受试者的研究。
- 在补充材料中包含此信息很好，但如果论文的主要贡献涉及人类受试者，则应包含尽可能多的细节。
- 根据NeurIPS 道德准则，参与数据收集、管理或其他劳动的工人应至少获得数据收集者所在国家/地区的最低工资。

#### 15. 人体研究的机构审查委员会 (IRB) 批准或同等批准

问题：论文是否描述了研究参与者产生的潜在风险，是否向受试者披露了此类风险，以及是否获得了机构审查委员会 (IRB) 的批准（或根据您所在国家或机构的要求进行的同等批准/审查）？

Answer: [NA]

理由：本文不涉及众包或人类受试者研究。

Guidelines:

- 答案NA 表示该论文不涉及众包或人类受试者研究。
- 根据进行研究的国家/地区，任何人类受试者研究可能需要IRB 批准（或同等文件）。如果您获得了 IRB 批准，您应该在文件中明确说明这一点。

- 我们认识到，不同机构和地点之间的程序可能会有很大差异，我们希望作者遵守 NeurIPS 道德准则及其机构指南。
- 对于初次提交的内容，请勿包含任何会破坏匿名性的信息（如果适用），例如进行审查的机构。