
Muse: Text-To-Image Generation via Masked Generative Transformers

Huiwen Chang^{*} Han Zhang^{*} Jarred Barber[†] AJ Maschinot[†] José Lezama Lu Jiang Ming-Hsuan Yang
Kevin Murphy William T. Freeman Michael Rubinstein[†] Yuanzhen Li[†] Dilip Krishnan[†]

Google Research

英寸

Abstract

我们展示了Muse，文本到图像的Transformer模型，实现了最先进的图像生成性能，同时比扩散或自回归模型更有效。Muse在离散令牌空间的蒙面建模任务上训练：给定从预训练的大型语言模型(LLM)中提取的文本嵌入，训练Muse来预测随机蒙面图像令牌。与像素空间扩散模型(如Imagen和DALL-E 2)相比，Muse由于使用离散令牌并且需要更少的采样迭代，因此效率显著更高；与自回归模型(如Parti)相比，Muse由于使用了并行解码，因此效率更高。使用预先训练的LLM可以实现细粒度的语言理解，翻译成高保真的图像生成，并理解视觉概念，如对象，它们的空间关系，姿态，cardinality等。我们的900M参数模型在CC3M上实现了新的SOTA，FID评分为6.06。Muse3B参数模型在零镜头COCO评估中实现了7.88的FID，以及0.32的CLIP评分。Muse还直接支持许多图像编辑应用程序，而不需要微调或反转模型：重新绘制、重新绘制和无遮罩编辑。更多结果请访问<http://muse-model.github.io>。

1. 简介

以文本提示为条件的生成图像模型在过去几年里在质量和灵活性方面取得了巨大的飞跃(Ramesh et al., 2022; Nichol et al., 2021; Saharia et al., 2022; Yu et al., 2022; Rombach et al., 2022; Midjourney, 2022)。这得益于深度学习架构创新的结合(Van Den Oord et al., 2017; Vaswani et al., 2017)；新颖的训练范式，如语言(Devlin et al., 2018; Raffel et al., 2020)和视觉任务的蒙面建模(He et al., 2022; Chang et al., 2022)；生成模型的新家族，如扩散(Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022)和基于掩码的生成(Chang et al., 2022)；最后，大规模图像-文本配对数据集的可用性(Schuhmann et al., 2021)。

在这项工作中，我们提出了一种新的文本到图像合成模型，使用掩码图像建模方法(Chang et al., 2022)。我们的图像解码器架构以预先训练和冻结的T5-XXL (Raffel et al., 2020)大型语言模型(LLM)编码器的嵌入为条件。与Imagen (Saharia et al., 2022)一致，我们发现预先训练的LLM对逼真的高质量图像生成至关重要。我们的模型(VQGAN量化器除外)构建在Transformer (Vaswani et al., 2017)体系结构上。

我们训练了一系列Muse模型，大小从632M参数到3B参数(用于图像解码器；T5-XXL型号有一个额外的4.6B参数)。每个模型由几个子模型组成(Figure 3)：首先，我们有一对VQGAN“标记器”模型(Esser et al., 2021b)，它将输入图像编码为离散标记序列，也可以将标记序列解码为图像。我们使用两个vqgan，一个用于256x256分辨率(“低分辨率”)，另一个用于512x512分辨率(“高分辨率”)。其次，我们有一个基本的蒙面图像模型，它包含了我们的大部分参数。该模型采用部分掩码的低分辨率令牌序列，并预测每个掩码令牌的边际分布，以未掩码令牌和T5XXL文本嵌入为条件。第三，我们有一个“superres”转换器模型，它将(未屏蔽的)低分辨率令牌转换为高分辨率令牌，再次以T5-XXL文本嵌入为条件。我们将在Section 2中详细解释我们的管道。

相比Imagen (Saharia et al., 2022)或dal - e2 (Ramesh et al., 2022)，建立在级联像素空间扩散模型上，Muse由于使用离散令牌，效率显著更高；它可以被认为是一个具有吸收状态的离散扩散过程([MASK])(Austin et al., 2021)。与最先进的自回归模型Parti (Yu et al., 2022)相比，Muse由于使用并行解码，效率更高。基于对类似硬件(TPU-v4芯片)的比较，我们估计Muse在推理时间上比image - 3b或partit - 3b模型快10 x以上，比Stable Diffusion v1.4 (Rombach et al., 2022)快3 x(参见Section 3.2.2)。所有这些比较都是在相同大小的图像(256×256 或 512×512)时进行的。Muse也比

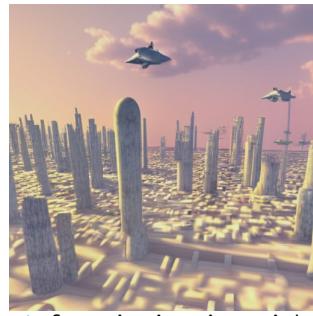
^{*}Equal contribution [†]Core contribution. Correspondence to: Huiwen Chang <huiwenchang@google.com>, Han Zhang <zhanghan@google.com>, Dilip Krishnan <dilipkay@google.com>.



A fluffy baby sloth with a knitted hat trying to figure out a laptop, close up.



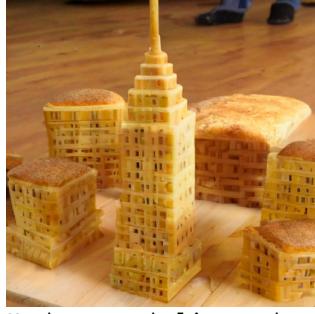
A sheep in a wine glass.



A futuristic city with flying cars.



A large array of colorful cupcakes, arranged on a maple table to spell MUSE.



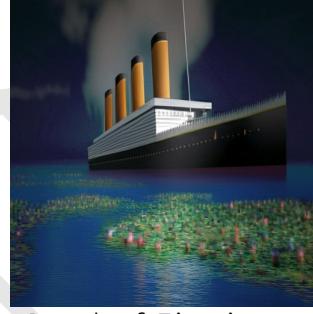
Manhattan skyline made of bread.



Astronauts kicking a football in front of Eiffel tower.



Two cats doing research.



3D mesh of Titanic floating on a water lily pond in the style of Monet.



A storefront with 'Muse' written on it, in front of Matterhorn Zermatt.



A surreal painting of a robot making coffee.



A cake made of macarons in a unicorn shape.



Three dogs celebrating Christmas with some champagne.

Figure 1. Muse文本到图像的生成(512×512 分辨率)。在每个生成的图像下，会显示相应的标题，展示出各种风格、标题和理解。每张图像都是在TPUv4芯片上的1.3 s中生成的。

稳定扩散(Rombach et al., 2022)快，尽管这两个模型都在VQGAN的潜在空间中工作。我们认为这是由于在Stable diffusion v1.4中使用了扩散模型，该模型在推理时需要大量的迭代。

但是，Muse的效率改进并不是以生成的图像质量或对输入文本提示的语义理解为代价的。我们根据多个标准评估我们的输出，包括CLIP评分(Radford et al., 2021)和FID (Heusel et al., 2017)。前者是图像-文本对应的度量；后者是衡量图像质量和多样性的标准。我们的3B参数模型在COCO (Lin et al., 2014)零镜头验证基准上获得了0.32的CLIP分数和7.88的FID分数，这与其他大规模文本到图像模型(参见Table 2)相比是很好的。我们的632M(基础)+268M(超分辨率)参数模型在CC3M (Sharma et al., 2018)数据集上训练和评估时，达到了6.06的FID评分，这明显低于文献中所有其他报道的结果(见Table 1)。我们还在PartiPrompts (Yu et al., 2022)评估套件上与人类评分员一起评估我们的代，他们发现Muse生成的图像与其文本提示2.7 x比Stable Diffusion v1.4 (Rombach et al., 2022)更好地匹配。

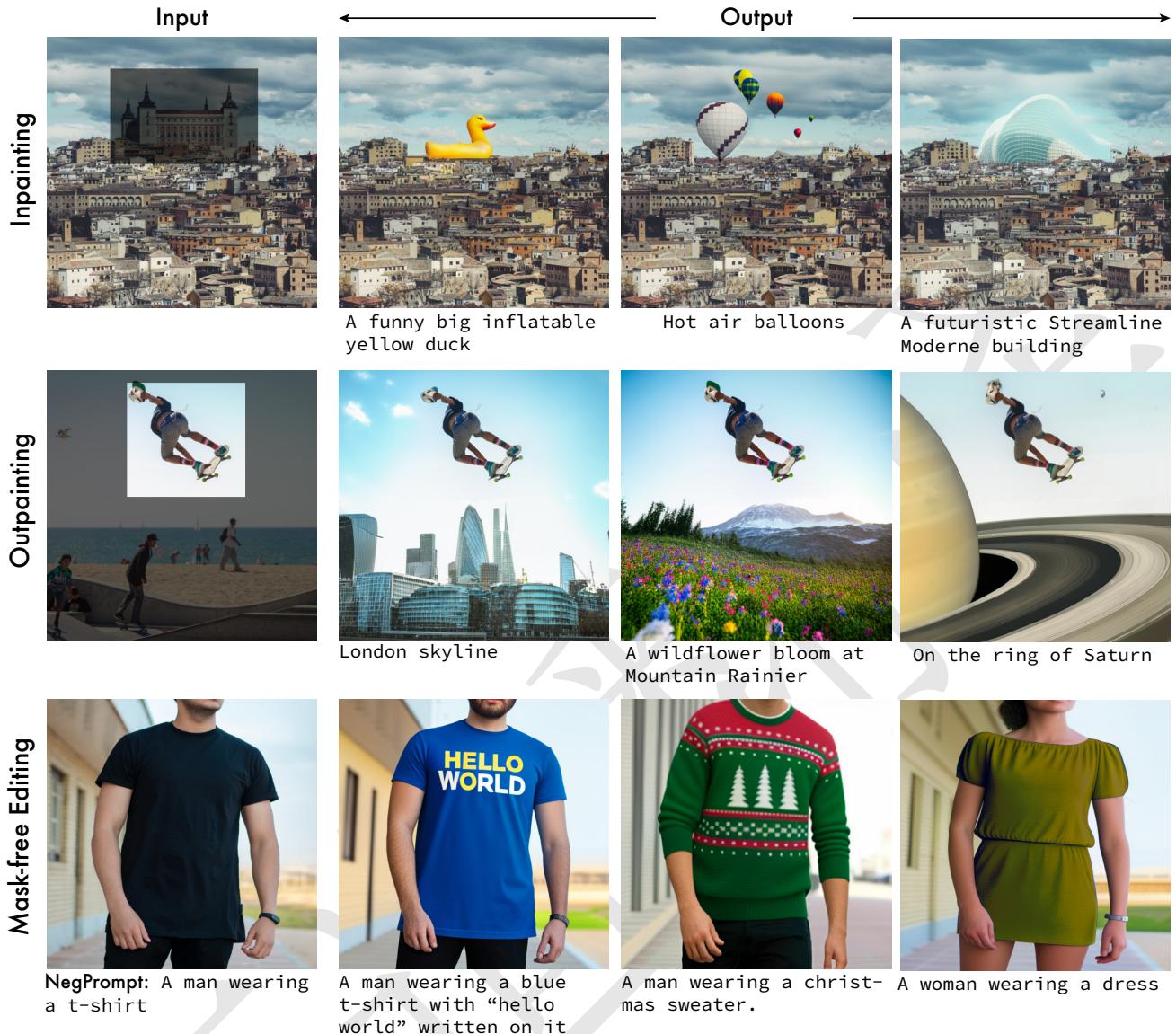


Figure 2. 使用Muse的零镜头文本引导图像编辑示例。我们展示了一些编辑应用程序的示例，这些应用程序使用Muse文本到图像的生成模型，在真实的输入图像上，没有进行微调。所有编辑的图像都是在 512×512 分辨率下生成的。

Muse生成图像，反映输入标题中的不同语音部分，包括名词、动词和形容词。此外，我们提出了多对象属性理解的证据，如组合性和基数，以及图像风格理解。更多示例请参见Figure 1，更多示例请参见我们的网站<http://muse-model.github.io>。Muse的基于掩模的训练使其具有许多零镜头图像编辑功能。Figure 2中展示了其中的一些功能，包括零拍摄、文本引导补绘、补绘和无遮罩编辑。更多详情见Section 3。我们的贡献是：

1. 我们提出了一个最先进的文本到图像生成模型，该模型获得了优秀的FID和CLIP分数(图像生成质量、多样性和与文本提示对齐的定量测量)。
2. 由于使用了量化图像标记和并行解码，我们的模型明显比可比模型快。
3. 我们的架构支持开箱即用的零镜头编辑功能，包括inpainting, outpainting和free mask编辑。

2. 模型

我们的模型是建立在许多组件之上的。在这里，我们按照训练的顺序对每个组件进行概述，而将架构和参数的

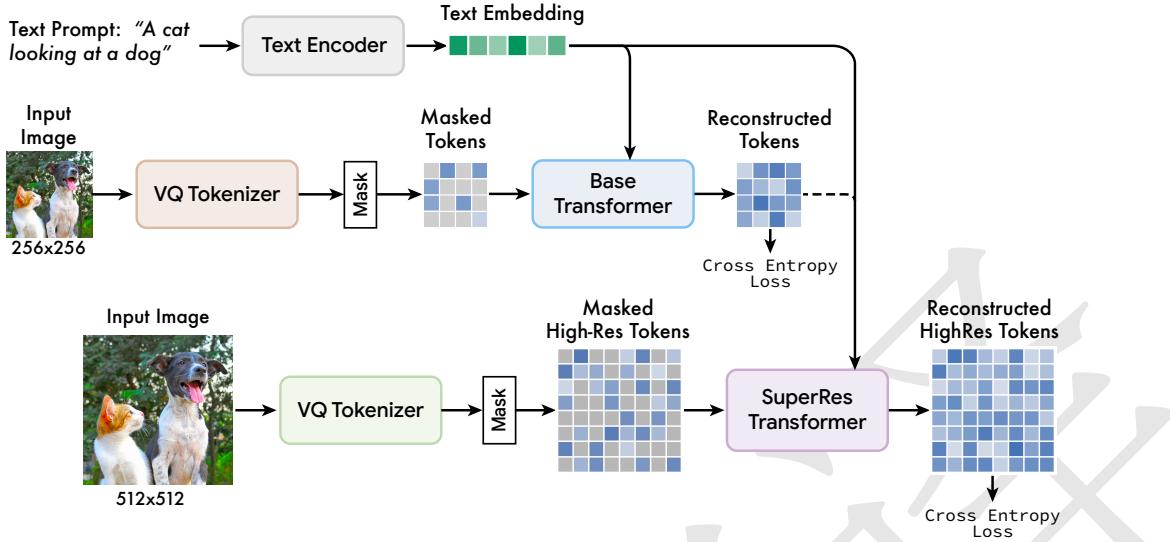


Figure 3. Muse框架:我们展示了模型的训练管道，T5-XXL预训练的文本编码器、基本模型和超分辨率模型描述在三行中。文本编码器生成文本嵌入，用于基本和超分辨率Transformer层的图像标记交叉注意。基本模型使用VQ Tokenizer，它在较低分辨率(256×256)图像上进行预训练，并生成 16×16 潜在的令牌空间。该序列以每个样本的可变速率被掩盖，然后交叉熵损失学习预测被掩盖的图像标记。一旦基本模型得到训练，重建的低分辨率令牌和文本令牌被传递到超分辨率模型，然后学习以更高分辨率预测掩码令牌。

许多细节放到附录中。Figure 3提供了模型体系结构的概述。

2.1. 预训练文本编码器

与(Saharia et al., 2022)中的发现类似，我们发现利用预训练的大型语言模型(LLM)有利于高质量的图像生成。从诸如T5-XXL (Raffel et al., 2020)这样的LLM中提取的嵌入包含关于对象(名词)、动作(动词)、视觉属性(形容词)、空间关系(介词)和其他属性(如基数和组合)的丰富信息。我们的假设是，Muse模型学会了将LLM嵌入中的这些丰富的视觉和语义概念映射到生成的图像；在最近的工作(Merullo et al., 2022)中已经表明，LLM学习的概念表示与视觉任务训练的模型学习的概念表示大致线性映射。给定一个输入文本标题，我们将其通过冻结的T5-XXL编码器，得到一个4096维语言嵌入向量序列。这些嵌入向量线性投影到Transformer模型的隐藏大小(基本和超分辨率)。

2.2. 使用VQGAN进行语义标记化

我们模型的一个核心组件是使用从VQGAN (Esser et al., 2021b)模型获得的语义标记。该模型由一个编码器和一个解码器组成，其中一个量化层将输入图像映射到一个学习码本中的令牌序列。我们完全使用卷积层构建编码器和解码器，以支持对不同分辨率的图像进行编码。编码器有几个下采样块来降低输入的空间维数，而解码器有相应数量的上采样块来将隐波映射回原始图像大小。给定一个大小为 $H \times W$ 的图像，编码的令牌的大小为 $H/f \times W/f$ ，下采样比为 f 。我们训练了两个VQGAN模型：一个是降采样比 $f = 16$ ，另一个是降采样比 $f = 8$ 。我们使用 $f = 16$ VQGAN模型在 256×256 像素图像上获取基本模型的令牌，从而得到空间大小为 16×16 的令牌。我们使用 $f = 8$ VQGAN模型在 512×512 图像上获得超分辨率模型的令牌，对应的令牌空间大小为 64×64 。正如之前的工作(Esser et al., 2021b)中提到的，编码后产生的离散令牌捕获图像的更高级别语义，同时忽略低级别噪声。此外，这些令牌的离散性质允许我们在下一阶段使用输出的交叉熵损失来预测掩码令牌。

2.3. 基本模型

我们的基本模型是一个屏蔽转换器(Vaswani et al., 2017; Devlin et al., 2018)，其中的输入是投影的T5嵌入和图像标记。我们保留所有的文本嵌入未蒙版，并随机蒙版不同比例的图像令牌(参见Section 2.6)，并用特殊的[MASK]令牌(Chang et al., 2022)替换它们。然后，我们将图像标记线性映射到所需Transformer输入/隐藏大小的图像输入嵌入以及学习的2D位置嵌入。遵循之前的变压器架构(Vaswani et al., 2017)，我们使用了几个变压器层，包括自注意块、交叉注意块和MLP块来提取特征。在输出层，使用MLP将每个掩码图像嵌入转换为一组logit(对应于VQGAN码本大小)，并以ground truth令牌标签为目标应用交叉熵损失。在训练中，训练基本模型来预测每一步的所有掩码令牌。然而，对于推断，掩码预测以迭代的方式执行，这显著提高了质量。详情见Section 2.8。

2.4. 超分辨率模型

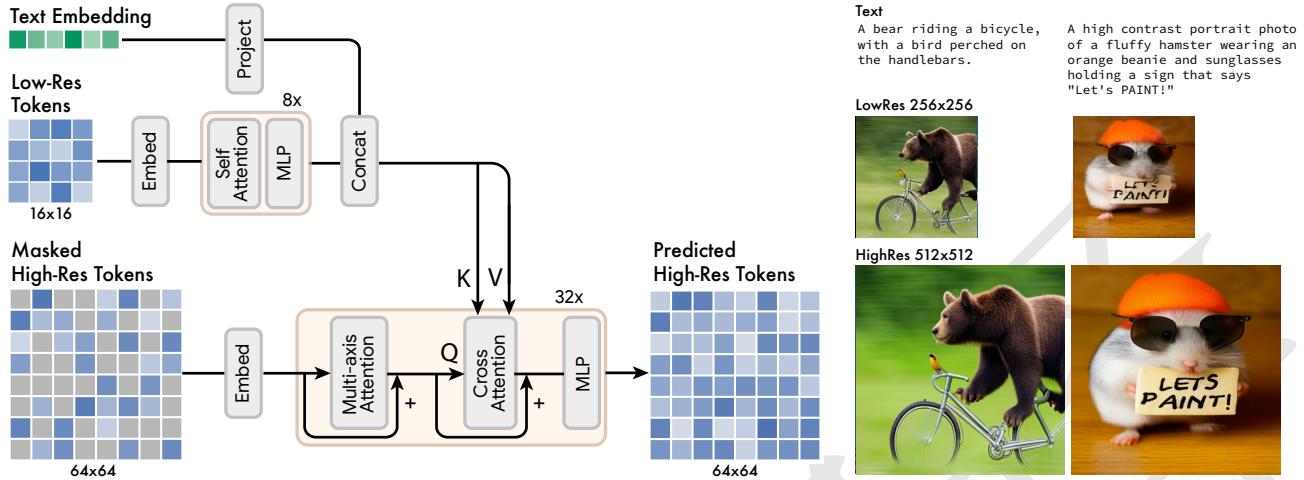


Figure 4. 超分辨率模型。左边是超分辨率模型的结构。低分辨率令牌被传递到一系列自关注Transformer层;结果输出嵌入与从条件文本提示符中提取的文本嵌入连接。在此之后，交叉注意力从这些连接的嵌入应用到被屏蔽的高分辨率令牌;在低分辨率和文本标记的条件下，该损失学会预测这些屏蔽标记。右边显示了超分辨率模型带来的改进的两个例子。

我们发现，直接预测 512×512 分辨率会导致模型关注低级细节而不是大规模语义。因此，我们发现使用级联模型是有益的:首先是一个基本模型，它生成 16×16 潜在映射(对应于 256×256 图像)，然后是一个超分辨率模型，它将基本潜在映射上采样到 64×64 潜在映射(对应于 512×512 图像)。超分辨率模型是在基本模型训练之后训练的。

正如Section 2.2中提到的，我们训练了两个VQGAN模型，一个是 16×16 潜在分辨率和 256×256 空间分辨率，另一个是 64×64 潜在分辨率和 512×512 空间分辨率。由于我们的基本模型输出对应于 16×16 潜在映射的令牌，我们的超分辨率过程学习将低分辨率潜在映射“翻译”为高分辨率潜在映射，然后通过高分辨率VQGAN解码，以给出最终的高分辨率图像。这个潜在的地图翻译模型也以类似于基础模型的方式用文本条件反射和交叉注意进行训练，如Figure 4所示。

2.5. 解码器微调

为了进一步提高我们的模型生成精细细节的能力，我们在保持编码器容量不变的情况下，通过添加更多残留层和通道来增加VQGAN解码器的容量。然后，我们对新的解码器层进行微调，同时保持VQGAN编码器权重、码本和变压器(即基本模型和超分辨率模型)不变。这允许我们在不重新训练任何其他模型组件的情况下提高我们的视觉质量(因为视觉标记“语言”保持固定)。这在附录中的Figure 13中显示，在那里我们看到经过微调的解码器可以在商店前面重建更清晰的细节。我们还在附录中给出了经过优化的解码器架构的细节。

2.6. 可变屏蔽率

正如在(Chang et al., 2022)中所做的那样，我们使用基于余弦调度的可变掩蔽率来训练我们的模型:对于每个训练示例，我们从截断的arccos分布中使用密度函数 $p(r) = \frac{2}{\pi}(1 - r^2)^{-\frac{1}{2}}$ 采样掩蔽率 $r \in [0, 1]$ 。它的预期掩蔽率为0.64，强烈倾向于更高的掩蔽率。对较高掩蔽率的偏向使得预测问题更加困难。与自回归方法相反，自回归方法为一些固定的令牌顺序学习条件分布 $P(x_i|x_{<i})$ ，具有可变屏蔽比的随机屏蔽允许我们的模型为令牌的任意子集 Λ 学习 $P(x_i|\Lambda)$ 。这不仅对我们的并行采样方案至关重要，而且还支持大量的零镜头、开箱即用的编辑功能，如Figure 2和Section 3.3所示。

2.7. 分类器自由引导

我们使用无分类器指导(CFG) (Ho & Salimans, 2022)来提高我们的生成质量和文本-图像对齐。在训练时，我们对随机选择的10%的样本去除文本条件(因此注意力减少到图像标记的自我注意)。在推理时，我们为每个掩码令牌计算一个条件logit ℓ_c 和一个无条件logit ℓ_u 。然后，我们形成最终的对数 ℓ_g ，从无条件对数中移出一个金额 t ，指导尺度:

$$\ell_g = (1 + t)\ell_c - t\ell_u \quad (1)$$

直观地说，CFG用多样性来换取保真性。与以前的方法不同，我们通过抽样过程线性增加指导尺度 t 来减少对多样性的打击。这允许更自由地对早期标记进行采样(很少或没有引导)，但增加了条件提示符对后期标记的影响。

我们还利用这种机制来启用消极提示(NegPrompt, 2022)，通过将无条件的logit ℓ_u 替换为以“消极提示”为条件的logit。这鼓励生成的图像具有与正面提示 ℓ_c 相关的特征，并删除与负面提示 ℓ_u 相关的特征。

2.8. 推理的迭代并行解码



Figure 5. 推断样本。我们在基本模型(左)和超分辨率模型(右)的步骤序列中可视化屏蔽标记的演变。超分辨率模型以低分辨率令牌为条件，需要更少的采样步骤进行收敛。

我们的模型推理时间效率的关键组件是使用并行解码来预测单个向前传递中的多个输出令牌。并行解码有效性的关键假设是一个马尔可夫性质，即许多令牌在给定其他令牌的情况下是有条件独立的。解码基于余弦计划(Chang et al., 2022)执行，该计划选择在该步骤预测的最高置信度掩码令牌的某个固定分数。然后将这些令牌设置为在其余步骤中取消掩码，并适当减少掩码令牌集。使用这个过程，我们能够在我们的基本模型中仅使用24解码步骤执行256令牌的推断，在我们的超分辨率模型中仅使用8解码步骤执行4096令牌的推断，相比之下，自回归模型(例如(Yu et al., 2022))需要256或4096步，扩散模型(例如(Rombach et al., 2022; Saharia et al., 2022))需要数百步。我们注意到最近的方法，包括渐进式蒸馏(Salimans & Ho, 2022)和更好的ODE求解器(Lu et al., 2022)，已经大大减少了扩散模型的采样步骤，但它们还没有在大规模文本到图像生成中得到广泛验证。在未来的工作中，我们把比较留给这些更快的方法，同时注意到类似的蒸馏方法也是我们模型的一种可能性。

3. 结果

我们训练了一些不同参数大小的基础Transformer模型，从600M到3B参数。这些模型中的每一个都来自T5-XXL模型的输出嵌入，该模型经过预训练和冻结，由4.6B参数组成。我们最大的3B参数基础模型由48 Transformer层组成，具有从文本到图像的交叉注意和图像令牌之间的自注意。所有基本模型共享相同的图像标记器。我们使用带有19 ResNet块的CNN模型和大小为8192的量化码本进行标记化。更大的码本大小并没有带来性能的提高。超分辨率模型由32多轴Transformer层(Zhao et al., 2021)组成，具有从拼接文本和图像嵌入到高分辨率图像的交叉注意以及高分辨率图像令牌之间的自注意。该模型将符号序列从一个潜在空间转换为另一个潜在空间：第一个潜在空间是基本模型标记器的潜在空间， 16×16 标记的潜在空间，到具有 64×64 标记的更高分辨率标记器的潜在空间。在令牌转换后，使用高分辨率令牌器的解码器将其转换为高分辨率图像空间。进一步的配置细节在附录中提供。

我们在由460M文本图像对(Saharia et al., 2022)组成的Imagen数据集上进行训练。训练为1M步，批次大小为512，使用512核TPU-v4芯片(Jouppi et al., 2020)。这需要大约1周的训练时间。我们使用Adafactor优化器(Shazeer & Stern, 2018)来节省内存消耗，这允许我们在没有模型并行化的情况下拟合一个3B参数模型。我们还避免在训练期间对模型权重执行指数移动平均(EMA)，这也是为了节省TPU内存。为了获得EMA的好处，我们每5000步检查一次，然后在衰减因子为0.7的检查点权重上离线执行EMA。这些平均权重形成了最终的基本模型权重。

3.1. 定性表现

Figure 6 定性地演示了Muse对于具有不同属性的文本提示的功能。Figure 6的左上角展示了一些示例，演示了对基数的基本理解。对于具有非统一基数的对象，Muse不是多次生成相同的对象像素，而是添加上下文变化以使整体图像更真实，例如，大象的大小和方向，酒瓶包装的颜色和网球旋转。图6的右上方展示了对多目标组成和相关性的理解。Muse生成的图像不是将对象放置在随机位置，而是保留文本中的介词对象关系，例如on vs under, left vs right等。Figure 6的中左显示了它生成跨越多种风格的图像的能力，既特定于著名艺术家(如伦勃朗)，也

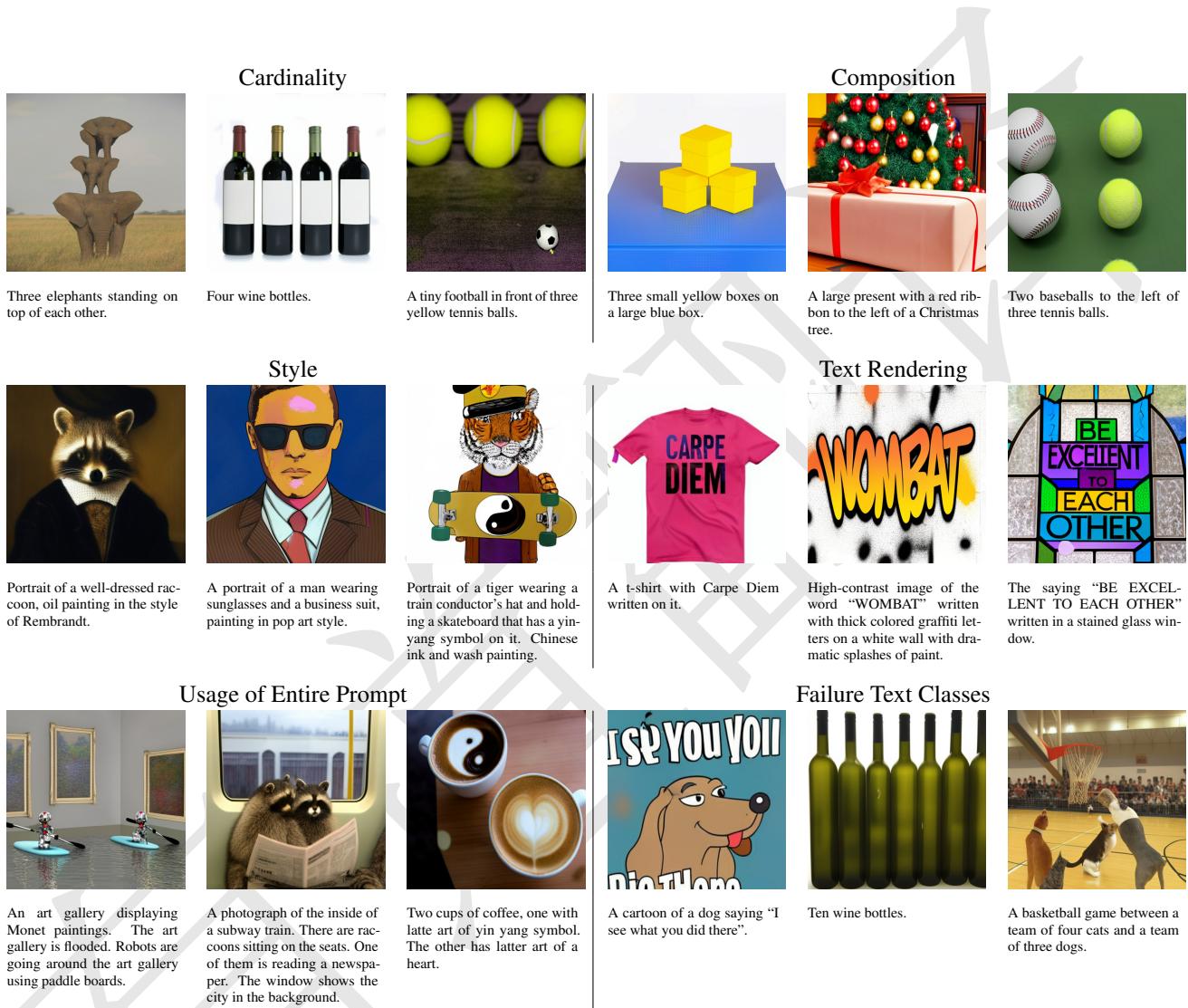


Figure 6. 示例演示了Muse的各种文本属性的文本到图像功能。左上:基数;右上:构图;左中:样式;右中:文本渲染;左下角:整个提示符的使用情况。对于所有示例，每个提示生成16个实例，并选择CLIP得分最高的实例(Radford et al., 2021)。右下:在Muse中生成的各种文本属性的图像失败示例，例如直接呈现长短语、高基数和多基数。



Figure 7. 在DALL-E2 (Ramesh et al., 2022)(左)、Imagen (Saharia et al., 2022)(中)和Muse(右)之间比较相同的提示。

Approach	Model Type	Params	FID	CLIP
VQGAN (Esser et al., 2021b)	Autoregressive	600M	28.86	0.20
ImageBART (Esser et al., 2021a)	Diffusion+Autoregressive	2.8B	22.61	0.23
LDM-4 (Rombach et al., 2022)	Diffusion	645M	17.01	0.24
RQ-Transformer (Lee et al., 2022a)	Autoregressive	654M	12.33	0.26
Draft-and-revise (Lee et al., 2022b)	Non-autoregressive	654M	9.65	0.26
Muse(base model)	Non-autoregressive	632M	6.8	0.25
Muse(base + super-res)	Non-autoregressive	632M + 268M	6.06	0.26

Table 1. CC3M定量评价(Sharma et al., 2018);所有模型都在CC3M上进行训练和评估。

Approach	Model Type	Params	FID-30K	Zero-shot FID-30K
AttnGAN (Xu et al., 2017)	GAN		35.49	-
DM-GAN (Zhu et al., 2019)	GAN		32.64	-
DF-GAN (Tao et al., 2020)	GAN		21.42	-
DM-GAN + CL (Ye et al., 2021)	GAN		20.79	-
XMC-GAN (Zhang et al., 2021)	GAN		9.33	-
LAFITE (Zhou et al., 2021)	GAN		8.12	-
Make-A-Scene (Gafni et al., 2022)	Autoregressive		7.55	-
DALL-E (Ramesh et al., 2021)	Autoregressive		-	17.89
LAFITE (Zhou et al., 2021)	GAN		-	26.94
LDM (Rombach et al., 2022)	Diffusion		-	12.63
GLIDE (Nichol et al., 2021)	Diffusion		-	12.24
DALL-E 2 (Ramesh et al., 2022)	Diffusion		-	10.39
Imagen-3.4B (Saharia et al., 2022)	Diffusion		-	7.27
Parti-3B (Yu et al., 2022)	Autoregressive		-	8.10
Parti-20B (Yu et al., 2022)	Autoregressive		3.22	7.23
Muse-3B	Non-Autoregressive		-	7.88

Table 2. 在MS-COCO (Lin et al., 2014)上对 256×256 图像分辨率进行FID和CLIP评分(如有)的定量评估。Muse的CLIP得分为0.32, 高于Imagen中报告的0.27。上表中的其他论文没有报告CLIP分数。

适用于整体风格(如波普艺术和中国水墨)。Figure 6的右中演示了Muse呈现单词和短语的能力。文本生成与生成大多数其他对象有本质上的不同。模型学习的不是对象名称与其特征之间的映射(例如, “大象”映射到“大”、“灰色”和“吃花生”), 而是可能的单词和短语的虚拟连续体要求模型以不同的方式学习。相反, 它必须学会在短语、单词和字母之间进行分层理解。Figure 6的左下角演示了Muse在呈现时使用整个文本提示, 而不是只关注几个突出的单词。最后, Figure 7显示了Muse、Dall-E2 (Ramesh et al., 2022)和Imagen (Saharia et al., 2022)之间的一些选择提示符的比较, 表明Muse与Imagen相当, 并且在许多提示符上质量优于Dall-E2。

然而, 正如Figure 6的右下角所示, Muse在生成与某些类型的提示对齐的图像方面能力有限。对于提示应该直接呈现长的多词短语的提示, Muse倾向于不正确地呈现这些短语, 经常导致(不必要的)重复呈现单词或只呈现短语的一部分。此外, 指示高对象基数的提示往往导致生成的图像不能正确反映所需的基数(例如, 当提示指定10时, 只呈现7酒瓶)。一般来说, Muse呈现正确的对象基数的能力会随着基数的增加而降低。Muse的另一个困难提示类型是具有多个基数的提示类型(例如, “四只猫和三只狗”)。对于这种情况, Muse在呈现时倾向于至少有一个基数不正确。

3.2. 量化表现

在Table 1和Table 2中, 我们通过Fréchet Inception Distance (FID) (Heusel et al., 2017)测量样本的质量和多样性, 以及CLIP (Radford et al., 2021)评分, 测量图像/文本对齐, 展示了我们在CC3M (Sharma et al., 2018)和COCO (Lin et al.,

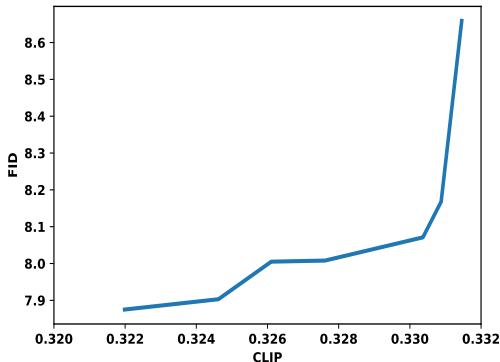


Figure 8. CLIP vs. FID权衡曲线。我们对固定模型的采样参数进行扫描，然后绘制帕累托前沿。

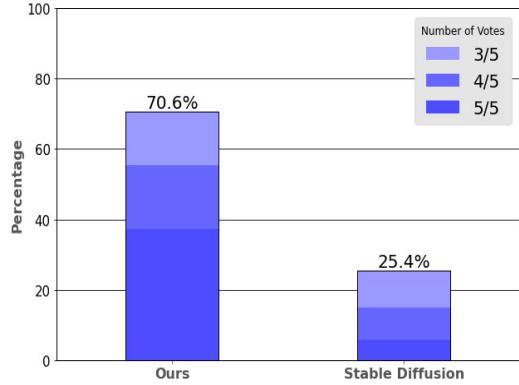


Figure 9. 人类评分者一致选择模型对齐偏好的提示的百分比。来自特定数量的评分者共识的贡献以不同的颜色显示，而共识的边缘值($=5$, ≥ 4 和 ≥ 3)以数字显示。

2014)数据集上与其他方法相比的性能。对于CC3M结果，两个Muse模型都在CC3M上进行了训练。COCO的结果是零拍摄的，使用了一个与Imagen (Saharia et al., 2022)在相同数据集上训练的模型。

我们的632M模型在CC3M上实现了SOTA结果，显著提高了FID评分的先进水平，也实现了CLIP评分的先进水平。我们的3B模型获得了7.88的FID分数，略好于拥有相似数量参数的Parti-3B模型获得的8.1的分数。我们0.32的CLIP得分高于Imagen的CLIP得分0.29(当FID显著高于20时实现)。对于7.27的FID, Imagen的CLIP得分约为0.27(参见(Saharia et al., 2022)中的图4)。

我们的采样算法(Section 2.8)有很多超参数，比如引导尺度、采样温度、采样过程中是否线性增加引导等。我们对这些参数执行评估扫描。我们发现抽样参数的子集是帕累托有效的，从某种意义上说，我们不能在不损害CLIP的情况下提高FID。这使我们能够研究多样性和图像/文本对齐之间的权衡，我们在Figure 8中展示了这一点。

3.2.1. 人的评价

类似于之前的工作(Yu et al., 2022; Saharia et al., 2022)，我们执行并排评估，其中人类评分者会看到一个文本提示和两张图像，每张图像都是由使用该提示的不同文本到图像模型生成的。评分者被要求通过“哪张图片与标题更匹配?”这个问题来评估提示图片的对齐程度。“每个图像对都是匿名和随机排序的(左与右)。评分者可以选择图像或他们无所谓¹。每个(提示，图像对)三元组由五个独立的评分者进行评估;评分者是通过谷歌内部人群计算团队提供的，对Muse团队完全匿名。对于呈现给评分者的提示集，我们使用partiprompt (Yu et al., 2022)，这是1650文本提示的集合，用于度量跨各种类别的模型功能。对于这两个文本到图像模型，我们将Muse(3 B参数)与Stable Diffusion v1.4 (Rombach et al., 2022)进行了比较，在推断速度方面，文本到图像模型与Muse最相似。对于每个提示，生成16图像实例，并使用CLIP得分最高的图像实例(Radford et al., 2021)。稳定扩散图像通过CompVis稳定扩散v1.4笔记本(CompVis, 2022)生成。我们至少需要3的评分者达成一致意见，才能计算有利于特定模型的结果。从这个分析中，我们发现对于70.6 %的提示，Muse被选择为比稳定扩散更好的对齐方式，对于25.4 %的提示，稳定扩散被选择为比Muse更好的对齐方式，而对于4 %的提示，没有选择评分者共识。这些结果与Muse一致，具有更好的标题匹配能力($\sim 2.7 \times$)。Figure 9显示了对3, 4和所有5可能投票的评分者一致意见的评分者结果的分解。所有5评分者认为Muse的对齐比稳定扩散更好的提示是更大的贡献者。

除了测量对齐之外，其他作品(Yu et al., 2022; Saharia et al., 2022)也测量了图像的真实感，通常通过类似于“哪个图像更真实?”的评分问题。然而，我们注意到，对这些结果的审查必须谨慎。虽然这不是这个问题的目的，但一个完全模式崩溃的模型，无论提示，它都会生成相同的足够逼真的图像，在这个问题上，实际上总是比一个做在图像生成过程中考虑提示的模型做得更好。我们认为这种类型的问题只适用于相似对齐的模型之间。由于Muse明显比稳定扩散更好，我们没有通过人类评分来评估真实感。我们认为这个主题是一个开放的研究领域。

¹当图像与文本提示都不一致时，选择无差异是有意义的，并有助于减少结果中的统计噪声。

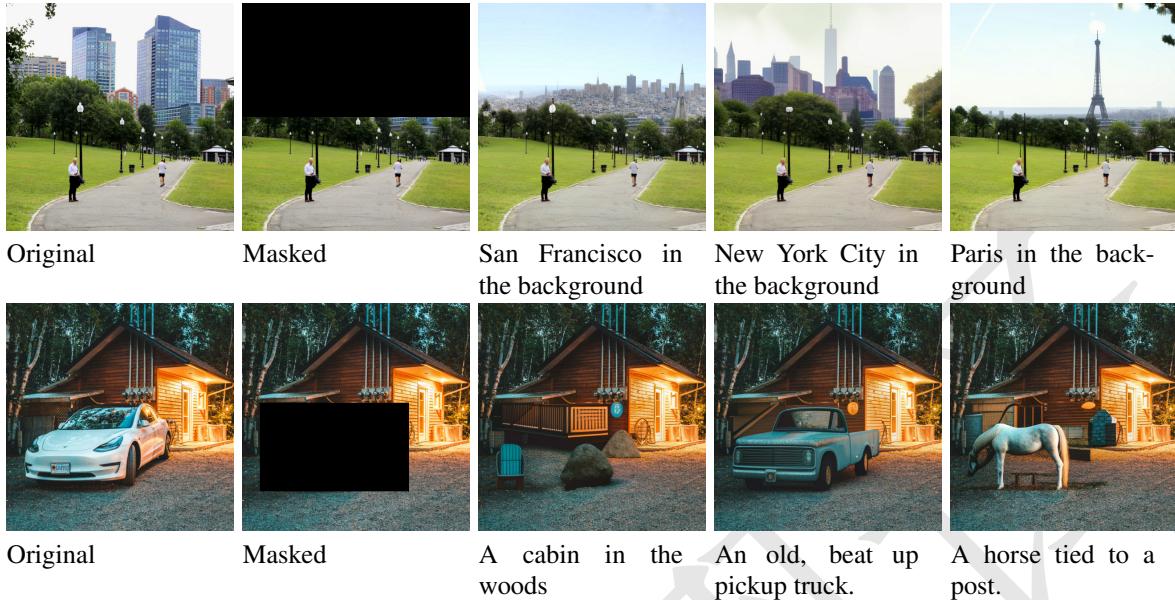


Figure 10. 文本引导修补的例子。掩码显示在每行的第二列中。这种行为直接来自模型，没有任何微调。

3.2.2. 推理速度

在Table 3中，我们将Muse的推断时间与其他几个流行模型进行比较。我们在TPUv4硬件上对Parti-3B、Imagen和Muse-3B进行了内部基准测试。对于稳定扩散/LDM，我们使用了最快的基准测试(Lambda Labs, 2022)，这是在A100 gpu上完成的。对于稳定扩散，我们测试的TPUv4实现了A100实现快51%。我们还报告了250次迭代的LDM的推理时间，这是用于在Table 2中实现FID的配置。Muse明显快于竞争扩散或自回归模型，尽管具有可比的参数数量(并且比稳定扩散/LDM多约3倍的参数)。Muse相对于Imagen的速度优势是由于使用了并行解码。Muse相对于稳定扩散的速度优势主要是由于需要更少的采样迭代。

Model	Resolution	Time
Imagen	256x256	9.1s
Parti-3B	512x512	13.3s
Muse-3B	256x256	1.3s

3.3. 图像编辑

通过利用我们的模型可以以图像标记的任意子集为条件的图像编辑。我们可以将模型用于50步的图像编辑应用程而无需额外的训练或模型微调。

3.3.1. 文本引导的补漆/出漆

我们的采样过程(Section 2.8)为我们免费提供了文本引导的补绘和出绘:我们将输入图像转换为一组标记，掩码出对应于局部区域的标记，然后根据未掩码标记和文本提示符对掩码标记进行采样。我们通过多尺度方法集成超分辨率:给定512x512大小的图像，我们首先将其抽取为256x256，并将这两个图像转换为高分辨率和低分辨率令牌。然后，我们为每一组标记屏蔽适当的区域。接下来，我们使用并行采样算法绘制低分辨率标记。最后，我们以这些低分辨率标记为条件，使用相同的采样算法重新绘制高分辨率标记。我们在Figure 2和Figure 10中展示了这样的例子。

3.3.2. 零镜头无遮罩编辑

我们使用Muse在零镜头意义上对真实输入图像进行无掩模图像编辑。这种方法直接作用于(标记化)图像，不需要“反转”完整的生成过程，与最近利用生成模型的零镜头图像编辑技术(Gal et al., 2022b; Patashnik et al., 2021; Kim et al., 2022; Mokady et al., 2022)形成对比。

我们首先将输入图像转换为可视标记。接下来，我们迭代地屏蔽和重新采样一个随机的标记子集，条件是文本提示。我们可以认为这类似于吉布斯抽样过程，我们固定一些令牌，并重新采样其他以它们为条件的令牌。这可以将标记化的图像移动到给定文本提示的图像的条件分布的典型集合中。

我们使用低分辨率的基本模型执行编辑，然后对最终输出执行超精度(以编辑提示符为条件)。在示例(Figure 2, Figure 11)中，我们对100次迭代中每次迭代重新采样8%的令牌，指导尺度为4。我们还在令牌日志上执行top- k ($k = 3$)采样，以防止进程与输入偏离太多。迭代特性允许对最终输出进行控制。Figure 12显示了一些中间编辑(没有超);在这个例子中，用户可能更喜欢迭代50或75而不是最终输出。

4. 相关工作

4.1. 图像生成模型

变分自编码器(Van Den Oord et al., 2017)和生成对抗模型(GANs)已经显示出出色的图像生成性能，为卷积和Transformer架构提出了许多变体，例如(Goodfellow et al., 2020; Esser et al., 2021b; Karras et al., 2019; Brock et al., 2018; Donahue & Simonyan, 2019)。直到最近，GANs都被认为是最先进的。基于渐进去噪原理的扩散模型现在能够以相同或更高保真度合成图像和视频(Ho et al., 2020; Kingma et al., 2021; Ho et al., 2022)。结合多种方法的原则的混合方法也表现出出色的性能(Chang et al., 2022; Lezama et al., 2022)，这表明可以利用的方法之间有更多的互补性。

4.2. 图像标记器

图像标记器被证明对多个生成模型是有用的，因为它能够将大量的计算从输入(像素)空间移动到潜在的(Rombach et al., 2022)，或者启用更有效的损失函数，如分类而不是回归(Chang et al., 2022; Lezama et al., 2022; Li et al., 2022)。已经开发了许多标记化方法，如Discrete VAE的(Rolfe, 2016), VQVAE (Van Den Oord et al., 2017)和VQGAN (Esser et al., 2021b)，后者是性能最高的，因为它结合了感知和对抗损失来实现出色的重建。vvi-VQGAN (Yu et al., 2021)将VQGAN扩展到Transformer体系结构。我们使用VQGAN而不是vvi-VQGAN，因为我们发现它对我们的模型性能更好，注意到性能更好的标记化模型并不总是转化为性能更好的文本到图像模型。

4.3. 大型语言模型

我们的工作利用T5，一个预先训练的大型语言模型(LLM)，已经在多个文本到文本任务上训练(Raffel et al., 2020)。llm(包括T5、BERT (Devlin et al., 2018)和GPT (Brown et al., 2020; Radford et al., 2019))已被证明可以学习强大的嵌入，从而实现少镜头迁移学习。我们在我们的模型中利用了这种能力。所有现代llm都是在标记预测任务(自回归或非自回归)上进行训练的。在这项工作中，我们利用了关于令牌预测功能的见解，其中我们应用了一个转换器来预测可视令牌。

4.4. 文本图像模型

利用配对文本图像数据被证明是表示学习和生成模型的强大学习范式。CLIP (Radford et al., 2021)和ALIGN (Jia et al., 2021)训练模型来对齐文本和图像嵌入对，显示出出色的传输和少镜头功能。Imagen (Saharia et al., 2022)和Parti (Yu et al., 2022)使用类似的大规模文本图像数据集(Schuhmann et al., 2021; 2022)来学习如何从文本输入中预测图像，在FID和人类评估上获得出色的结果。一个关键的技巧是使用无分类器的指导(Ho & Salimans, 2022; Dhariwal & Nichol, 2021)，这平衡了多样性和质量。

4.5. 用生成模型编辑图像

GANs在图像编辑和操作能力方面已被广泛研究(有关调查，请参阅(Xia et al., 2022))。在扩散模型上开发了许多技术，以实现对令牌空间的编辑、个性化和反转(Gal et al., 2022a; Meng et al., 2021; Ruiz et al., 2022; Kawar et al., 2022; Brooks et al., 2022; Hertz et al., 2022; Mokady et al., 2022)。Dreambooth (Ruiz et al., 2022)和image (Kawar et al., 2022)涉及对生成模型的微调。ImagenEditor (Wang et al., 2022)将编辑任务框成文本引导的图像修补，并涉及用户指定的蒙版。

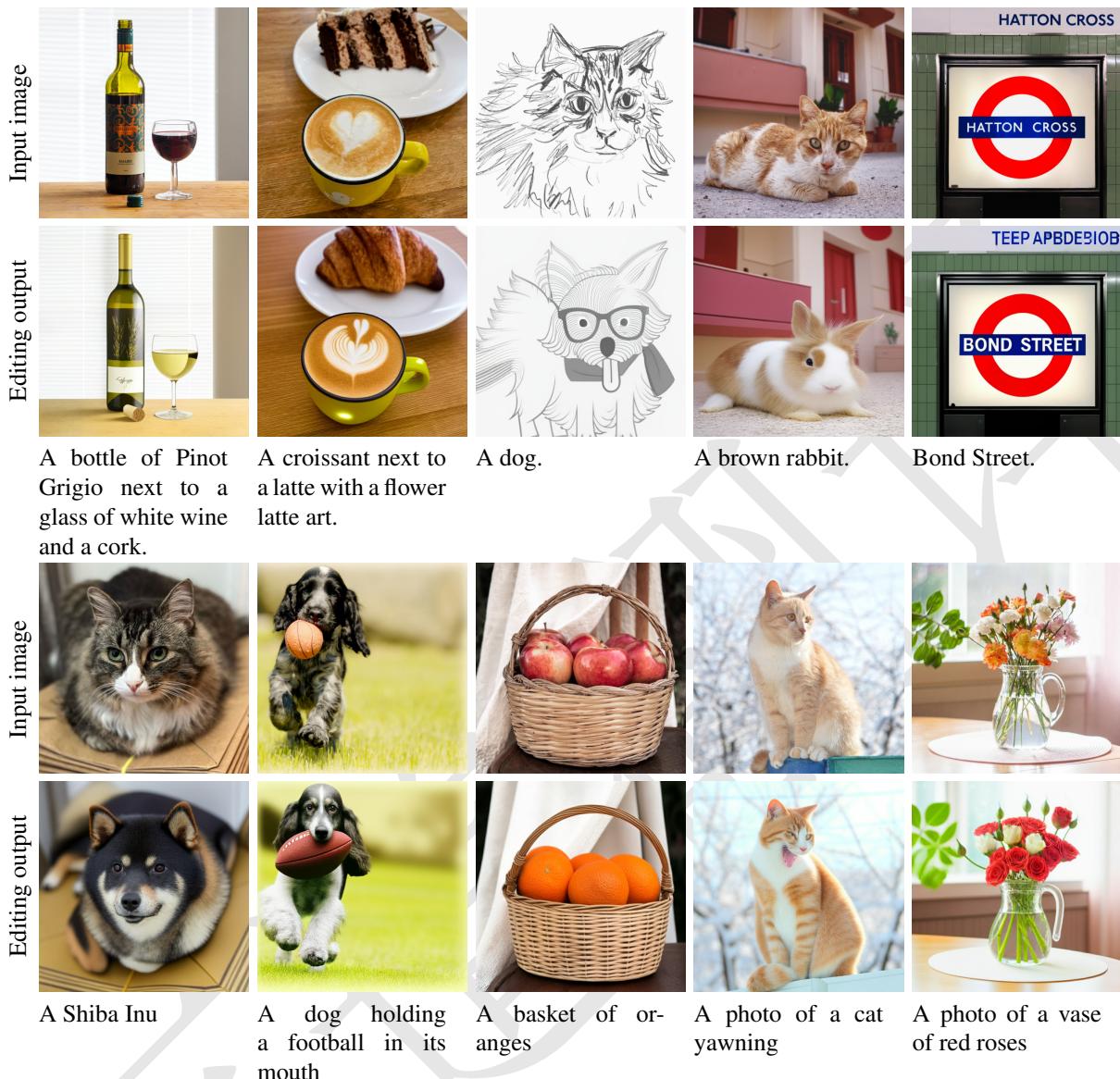


Figure 11. 零镜头无遮罩图像编辑的例子, post superres。我们看到, 图像的姿态和整体结构保持不变, 同时根据文本提示改变对象的某些特定方面。



Figure 12. 在Figure 11中产生一个编辑的中间迭代(pre-superres)

5. 讨论与社会影响

Muse模型证实了(Saharia et al., 2022)的发现，即冻结的大型预训练语言模型可以作为强大的文本编码器文本到图像的生成。在最初的实验中，我们还尝试从训练数据中从头开始学习语言模型，但发现性能明显不如使用预训练的LLM，特别是在长提示和罕见单词上。我们还表明，基于Transformer架构的非扩散、非自回归模型可以与扩散模型表现相当，同时在推理时效率显著更高。我们实现了SOTA CLIP分数，显示了图像和文本之间的优秀对齐。我们还通过一些图像编辑应用程序展示了我们的方法的灵活性。

我们认识到生成模型有许多应用程序，对人类社会有不同的潜在影响。生成模型(Saharia et al., 2022; Yu et al., 2022; Rombach et al., 2022; Midjourney, 2022)拥有巨大的潜力来增强人类的创造力(Hughes et al., 2021)。然而，众所周知，它们也可以用来制造错误信息、骚扰和各种类型的社会和文化偏见(Franks & Waldman, 2018; Whittaker et al., 2020; Srinivasan & Uchino, 2021; Steed & Caliskan, 2021)。由于这些重要的考虑，我们选择不在此时发布代码或公开演示。

数据集偏差是另一个重要的道德考虑因素，因为大型数据集大多是自动管理的。这样的数据集有各种潜在的问题，如同意和主体意识(Paullada et al., 2021; Dulhanty, 2020; Scheuerman et al., 2021)。许多常用的数据集往往反映负面的社会刻板印象和观点(Prabhu & Birhane, 2020)。因此，在这样的数据集上进行训练只会放大这些偏差是非常可行的，需要对如何减轻这些偏差进行重要的额外研究，并生成没有这些偏差的数据集：这是一个非常重要的主题(Buolamwini & Gebru, 2018; Hendricks et al., 2018)，但超出了本文的范围。

鉴于上述考虑，我们不建议在不注意各种用例和了解潜在危害的情况下使用文本到图像生成模型。我们特别警告不要将这种模型用于人、人和面孔的生成。

鸣谢

我们感谢William Chan、Chitwan sahara和Mohammad Norouzi为我们提供训练数据集、各种评估代码和慷慨的建议。Jay Yagnik, Rahul Sukthankar, Tom Duerig和David Salesin为这个项目提供了热情的支持，我们对此表示感谢。我们感谢Victor Gomes和Erica Moreira对基础设施的支持，Jing Yu Koh和Jason Baldridge对数据集、模型和评估的讨论和对论文的反馈，Mike Krainin对模型加速的讨论，JD Velasquez的讨论和见解，Sarah Laszlo、Kathy Meier-Hellstern和Rachel Stigler对我们发表过程的协助，Andrew Bunner、Jordi Pont-Tuset和Shai Noy对内部演示的帮助，David Fleet、Saurabh Saxena、Jiahui Yu，Jason Baldridge分享Imagen和Parti速度指标。

References

- Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.
- Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners.

Advances in neural information processing systems, 33: 1877–1901, 2020.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

CompVis. Stable diffusion colab, 2022. URL https://colab.sandbox.google.com/github/huggingface/notebooks/blob/main/diffusers/stable_diffusion.ipynb#scrollTo=zHkHsdtnry57.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Donahue, J. and Simonyan, K. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019.

Dulhanty, C. Issues in computer vision data collection: Bias, consent, and label taxonomy. Master’s thesis, University of Waterloo, 2020.

Esser, P., Rombach, R., Blattmann, A., and Ommer, B. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34:3518–3532, 2021a.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021b.

Franks, M. A. and Waldman, A. E. Sex, lies, and videotape: Deep fakes and free speech delusions. *Md. L. Rev.*, 78: 892, 2018.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. URL <https://arxiv.org/abs/2203.13131>.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation

using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022a.

Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., and Cohen-Or, D. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022b.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *preprint arXiv:1706.0267*, 2017.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *cvpr*, pp. 16000–16009, June 2022.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 771–787, 2018.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Hughes, R. T., Zhu, L., and Bednarz, T. Generative adversarial networks–enabled human–artificial intelligence collaborative applications for creative and design industries: A systematic review of current approaches and trends. *Frontiers in artificial intelligence*, 4:604234, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with

noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Jouppi, N. P., Yoon, D. H., Kurian, G., Li, S., Patil, N., Laudon, J., Young, C., and Patterson, D. A domain-specific supercomputer for training deep neural networks. *Communications of the ACM*, 63(7):67–78, 2020.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseli, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

Kim, G., Kwon, T., and Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Lambda Labs. All you need is one gpu: Inference benchmark for stable diffusion, 2022. URL <https://lambdalabs.com/blog/inference-benchmark-stable-diffusion>.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022a.

Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. Draft-and-revise: Effective image generation with contextual rq-transformer. *arXiv preprint arXiv:2206.04452*, 2022b.

Lezama, J., Chang, H., Jiang, L., and Essa, I. Improved masked image generation with token-critic. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2022.

Li, T., Chang, H., Mishra, S. K., Zhang, H., Katahi, D., and Krishnan, D. Mage: Masked generative encoder to unify representation learning and image synthesis. *arXiv preprint arXiv:2211.09117*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

- Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *iclr*, 2017.
- Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.
- Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.
- Midjourney. Midjourney, 2022. URL <https://www.midjourney.com>.
- Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models, 2022. URL <https://arxiv.org/abs/2211.09794>.
- NegPrompt. Negative prompt, 2022. URL <https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt>.
- Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.
- Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.
- Prabhu, V. U. and Birhane, A. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural

language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation, 2021. URL <https://arxiv.org/abs/2102.12092>.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rolfe, J. T. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Scheuerman, M. K., Hanna, A., and Denton, E. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Srinivasan, R. and Uchino, K. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 41–51, 2021.

Steed, R. and Caliskan, A. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 701–713, 2021.

Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K., and Xu, C. Df-gan: A simple and effective baseline for text-to-image synthesis, 2020. URL <https://arxiv.org/abs/2008.05865>.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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

Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D. J., Soricut, R., Baldridge, J., Norouzi, M., Anderson, P., and Chan, W. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting, 2022. URL <https://arxiv.org/abs/2212.06909>.

Whittaker, L., Kietzmann, T. C., Kietzmann, J., and Dabirian, A. “all around me are synthetic faces”: the mad world of ai-generated media. *IT Professional*, 22(5): 90–99, 2020.

Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., and Yang, M.-H. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017. URL <http://arxiv.org/abs/1711.10485>.

- Ye, H., Yang, X., Takac, M., Sunderraman, R., and Ji, S. Improving text-to-image synthesis using contrastive learning, 2021. URL <https://arxiv.org/abs/2107.02423>.
- Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.
- Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.
- Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.
- Zhao, L., Zhang, Z., Chen, T., Metaxas, D. N., and Zhang, H. Improved transformer for high-resolution gans, 2021. URL <https://arxiv.org/abs/2106.07631>.
- Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., and Sun, T. LAFITE: towards language-free training for text-to-image generation. *CoRR*, abs/2111.13792, 2021. URL <https://arxiv.org/abs/2111.13792>.
- Zhu, M., Pan, P., Chen, W., and Yang, Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5810, 2019.

A. 附录。

A.1. 基本模型配置

我们最大尺寸3B参数模型的基本模型配置在Table 4中给出。

Configuration	Value
Number of Transformer layers	48
Transformer Hidden Dimension	2048
Transformer MLP Dimension	8192
Optimizer	AdaFactor (Shazeer & Stern, 2018)
Base learning rate	1e-4
Weight decay	0.045
Optimizer momentum	$\beta_1=0.9, \beta_2=0.96$
Batch size	512
Learning rate schedule	cosine decay (Loshchilov & Hutter, 2017)
Warmup steps	5000
Training steps	1.5M

Table 4. 基本模型的配置和训练超参数。

A.2. VQGAN配置

Configuration	Value
Perceptual loss weight	0.05
Adversarial loss weight	0.1
Codebook size	8192
Optimizer	Adam (Kingma & Ba, 2015)
Discriminator learning rate	1e-4
Generator learning rate	1e-4
Weight decay	1e-4
Optimizer momentum	$\beta_1=0.9, \beta_2=0.99$
Batch size	256
Learning rate schedule	cosine decay (Loshchilov & Hutter, 2017)
Warmup steps (Goyal et al., 2017)	10000
Training steps	1M

Table 5. VQGAN的配置和训练超参数。

VQGAN架构:我们的VQGAN架构类似于以前的工作(Esser et al., 2021b)。它由几个残差块、下采样(编码器)块和上采样(解码器)块组成。主要的区别是，我们删除了非局部块，使编码器和解码器完全卷积，以支持不同的图像大小。在基本VQGAN模型中，我们在每个分辨率中应用2个残差块，基本信道维度为128。对于经过微调的解码器，我们在每个分辨率中应用4个残差块，我们也使基信道维度为256。



Figure 13. 来自微调解码器的改进的可视化示例(Section 2.5)。请放大至少200%，以查看VQGAN重建和经过微调的解码器重建之间的差异。我们尤其可以看到，在经过微调的解码器中，门牌号(左下)、店面标志(中)和窗户上的横条(右)等细节得到了更好的保存。

A.3. 超分辨率配置

Configuration	Value
LowRes Encoder Transformer Layers	16
Number of Transformer layers	32
Transformer Hidden Dimension	1024
Transformer MLP Dimension	4096
Optimizer	AdaFactor (Shazeer & Stern, 2018)
Base learning rate	1e-4
Weight decay	0.045
Optimizer momentum	$\beta_1=0.9, \beta_2=0.96$
Batch size	512
Learning rate schedule	cosine decay (Loshchilov & Hutter, 2017)
Warmup steps	5000
Training steps	1M

Table 6. 超分辨率模型的配置和训练超参数。