arXiv:2301.00704v1 [cs.CV] 2 Jan 2023

Muse：通过掩码生成式 Transformer 实现文本到图像生成

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

Google Research

摘要

我们介绍了 Muse，这是一种文本到图像 Transformer 模型，在实现最先进的图像生成性能的同时，效率显著高于扩散模型或自回归模型。Muse 在离散词元空间中通过掩码建模任务进行训练：给定从预训练大语言模型（LLM）中提取的文本嵌入，Muse 经过训练，可预测随机被掩码的图像令牌。与 Imagen 和 DALL-E 2 等像素空间扩散模型相比，Muse 得益于离散令牌的使用且所需采样迭代次数更少，因此效率显著更高；与 Parti 等自回归模型相比，Muse 则因采用并行解码而更加高效。预训练大语言模型的使用赋予了 Muse 细粒度的语言理解能力，从而实现高保真图像生成，并能理解物体、物体之间的空间关系、姿势、基数等视觉概念。我们的 900M 参数模型在 CC3M 上取得了新的最先进水平，FID 分数为 6.06。Muse 3B 参数模型在 COCO 零样本评估中取得了 7.88 的 FID 分数，CLIP 分数为 0.32。Muse 还可直接支持多种图像编辑应用，无需微调或反演模型：图像内补、图像外绘和无掩码编辑。更多结果请见http://muse-model.github.io。

1. 引言

过去几年里，以文本提示为条件的生成式图像模型在质量和灵活性方面取得了巨大飞跃（Ramesh 等人，2022年；Nichol 等人，2021年；Saharia 等人，2022年；Yu 等人，2022年；Rombach 等人，2022年；Midjourney，2022年）。这一进步得益于多方面因素的共同作用：深度学习架构的创新（VanDenOord 等人，2017年；Vaswani 等人，2017年）；新的训练范式，例如用于语言（Devlin 等人，2018；Raffel 等，2020年）和视觉任务（He 等人，2022年；Chang 等，2022）的掩码建模；扩散（Ho 等人，2020年；Rombach 等人，2022年；Saharia 等人，2022年）和基于掩码的生成（Chang 等，2022）等新型生成模型；以及大规模图文配对数据集的问世（Schuhmann 等人，2021年）。

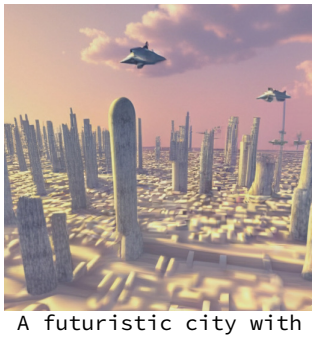
在这项工作中，我们采用掩码图像建模方法，提出了一种新的文本到图像合成模型（Chang 等人，2022年）。我们的图像解码器架构以预训练且冻结的 T5-XXL（Raffel 等人，2020年大语言模型（LLM）编码器生成的嵌入为条件。与 Imagen（Saharia 等人，2022年）的结果一致，我们发现，以预训练的大语言模型为条件，对于生成逼真、高质量的图像至关重要。我们的模型（VQGAN 量化器除外）均基于 Transformer（Vaswani 等人，2017年）架构构建。

我们训练了一系列 Muse 模型，参数规模从 6.32 亿到 30 亿参数不等（图像解码器的参数量；T5-XXL 模型还额外包含 46 亿参数）。每个模型由若干子模型组成（图3）：首先，我们使用一对 VQGAN“分词器”模型（Esser 等人，2021b），它们既能将输入图像编码为一串离散令牌，也能将令牌序列解码回图像。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，我们使用一个基础掩码图像模型，其中包含了我们的大部分参数。该模型接收一串部分被掩码的低分辨率令牌，并预测边缘分布

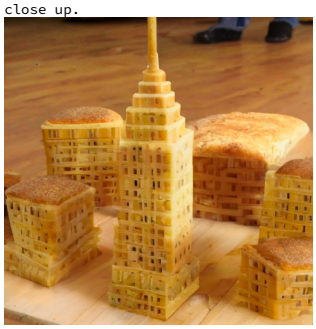
[*]贡献相同 [†]核心贡献。联系：Huiwen Chang <huiwenchang@google.com>，Han Zhang <zhang- han@google.com>，Dilip Krishnan <dilipkay@google.com>。















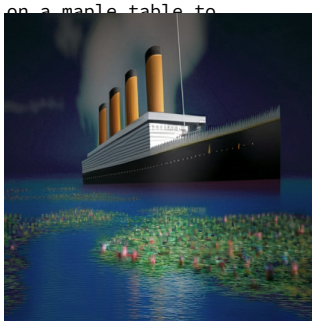










图*1*。Muse 文本到图像 生成（512 × 512 分辨率）。每张生成图像下方均显示对应的说明文字 ，
展现出多种风格， 图像说明文字与理解。每张图像均由 TPUv4 芯片在 1.3 秒内生成。

对每个掩码词元，根据未掩码词元和 T5XXL 文本嵌入进行条件生成。第三，我们使用一个“超分辨率”Transformer 模型，将（未掩码的）低分辨率词元转换为高分辨率词元，同样以 T5-XXL 文本嵌入为条件。我们将在第 2 节详细介绍我们的流程。

与 Imagen（Saharia 等人，2022年）或 Dall-E2（Ramesh 等人，2022年）相比，后者基于级联像素空间扩散模型构建，Muse 由于使用离散词元而显著更加高效；可以将其视为一个带有吸收态（[掩码]）的离散扩散过程（Austin 等人，2021年）。与 Parti（Yu 等人，2022年）这一先进的自回归模型相比，Muse 得益于并行解码而更加高效。基于相似硬件（TPU-v4 芯片）上的比较，我们估计，在推理时间方面，Muse 比 Imagen-3B 或 Parti-3B 模型快 10 倍以上，比 Stable Diffusion v1.4（Rombach 等人，2022年）（见第 3.2.2 节）快 10 倍以上，并且比 Stable Diffusion v1.4 快 3 倍。所有这些比较均基于相同尺寸的图像：即 256×256 或 512×512。尽管两种模型都在 VQGAN 的潜在空间中运行，Muse 仍然比 Stable Diffusion（Rombach 等人，2022年）更快。我们认为，这是由于 Stable

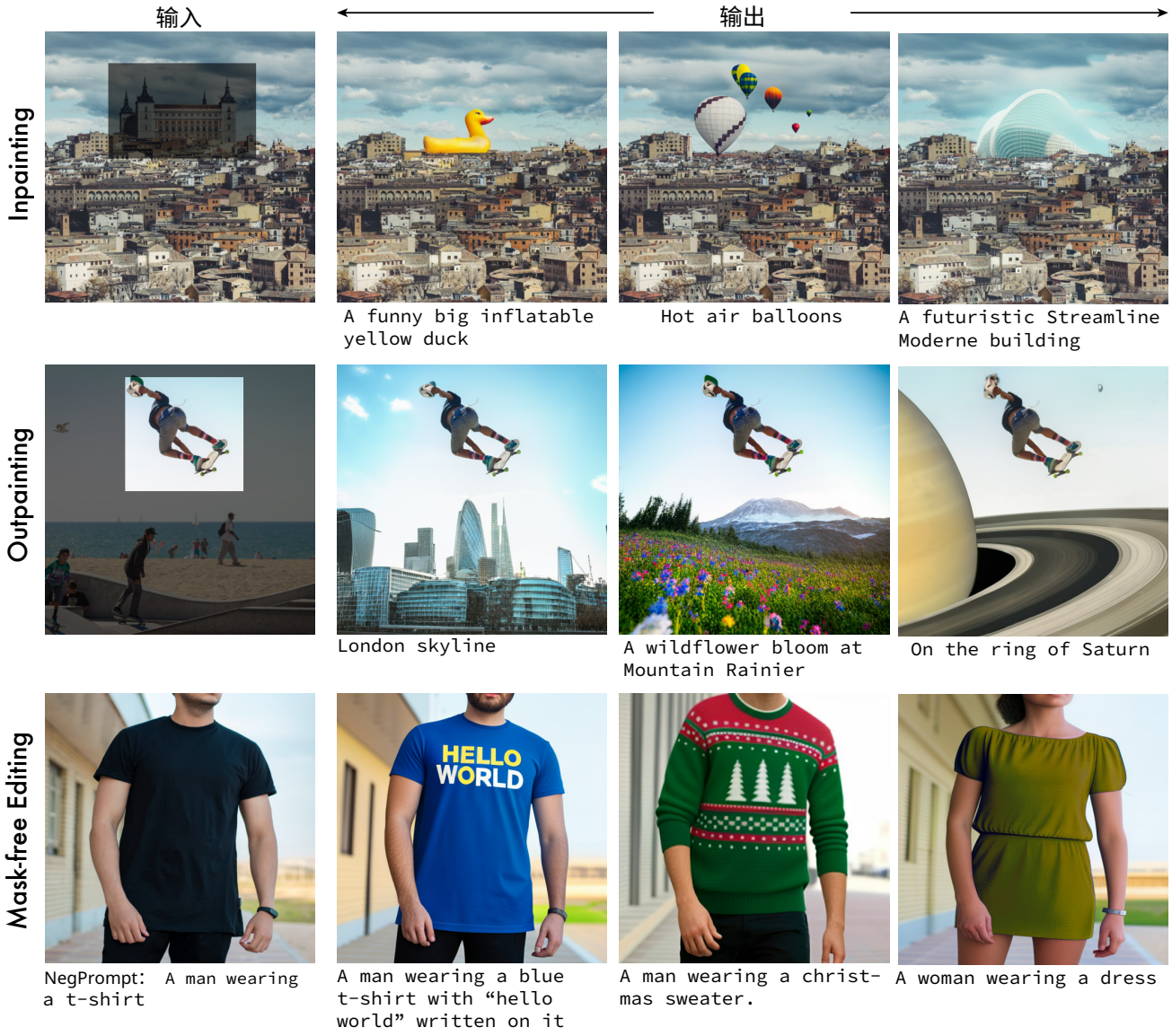


图2。 零样本 使用 Muse 进行文本引导的图像编辑。我们展示了使用该模型进行多种编辑应用的示例。 e
Muse 文本到图像生成 模型在真实输入图像上无需微调。所有编辑后的图像均以 512×512 分辨率生成 .

Diffusion v1.4 在推理时需要显著更多的迭代次数。

然而，Muse 的效率提升并未以牺牲生成图像的质量或对输入文本提示的语义理解为代价。我们根据多项标准评估输出，包括 CLIP 分数（Radford 等人，2021年）和 FID（Heusel 等人，2017年）。前者用于衡量图文对应关系，后者用于衡量图像质量与多样性。我们的 3B 参数模型在 COCO（Lin 等人，2014年）零样本验证基准上取得了 0.32 的 CLIP 分数和 7.88 的 FID 分数，相较于其他大规模文生图模型表现出色（见表 2）。我们的 632M（基础模型）+268M（超分辨率版）参数模型在 CC3M（Sharma 等人，2018）数据集上训练和评估时，取得了 6.06 的最先进 FID 分数，显著低于文献中报告的所有其他结果（见表 1）。我们还使用人工评审员在 PartiPrompts（Yu 等人，2022年）评估套件上评估生成结果。评审员认为，Muse 生成的图像与文本提示的契合度高于 Stable Diffusion v1.4（Rombach 等人，2022年）2.7倍的情况更常见。

Muse 能够生成图像，呈现输入图像说明文字中不同词性的内容，包括名词、动词和形容词。此外，我们还展示了模型对多物体属性的理解能力，例如组合性和基数，以及

良好的图像风格理解。若要查看其中一些示例，请参见图1；更多示例请访问我们的网站 http://muse-model.github.io。Muse 的基于掩码的训练使其具备多种零样本图像编辑能力。图2展示了其中一些，包括零样本文本引导的图像内绘制、图像外绘和无掩码编辑。更多详情见第3节。我们的贡献如下：

1. 我们提出了一种先进的文本到图像生成模型，在 FID 和 CLIP 分数方面表现出色（这些指标用于量化图像生成质量、多样性以及与文本提示的对齐程度）。2. 由于采用了量化图像令牌和并行解码，我们的模型比同类模型快得多。3. 我们的架构开箱即用，具备零样本编辑能力，包括图像内补、图像外绘和无掩码编辑。

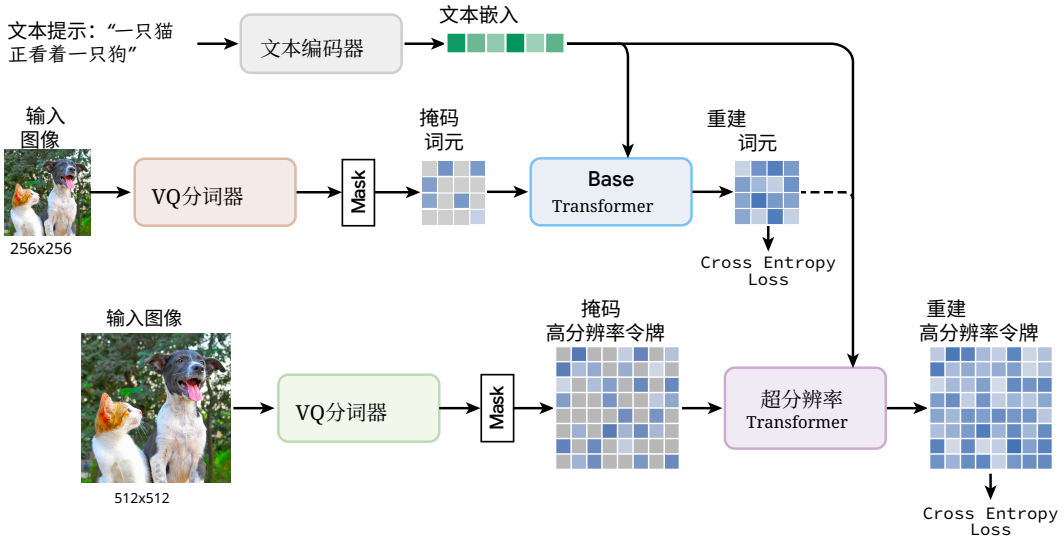


图 *3.* Muse框架：我们展示了模型的训练流程，三行分别描绘了 T5-XXL 预训练文本编码器、基础模型和超分辨率模型。文本编码器生成文本嵌入，用于基础模型和超分辨率Transformer层中与图像令牌进行交叉注意力。基础模型使用在较低分辨率（256×256）图像上预训练的 VQ分词器，并生成由令牌组成的 16×16 潜在空间。该序列会针对每个样本以不同的比例进行掩码，然后通过交叉熵损失学习预测被掩码的图像令牌。基础模型训练完成后，重建的低分辨率令牌和文本令牌会被输入超分辨率模型，后者随后学习预测更高分辨率下的掩码令牌。

2. 模型

我们的模型由多个组件构成。这里，我们按照这些组件的训练顺序逐一概述，同时将架构和参数的许多细节留待附录说明。图 3 展示了模型架构概览。

2.1. 预训练文本编码器

与（Saharia 等人，2022年）中的发现类似，我们发现，利用预训练大语言模型（LLM）有助于生成高质量图像。从 T5-XXL（Raffel 等人，2020年）等大语言模型中提取的嵌入，包含有关物体（名词）、动作（动词）、视觉属性（形容词）、空间关系（介词）以及基数和构图等其他属性的丰富信息。我们的假设是，Muse 模型能够学习将大语言模型嵌入中的丰富视觉和语义概念映射到生成的图像上；近期研究（Merullo 等人，2022年）表明，大语言模型学习到的概念表征与视觉任务模型学习到的概念表征大致可以通过线性映射相互转换。给定输入文本描述，我们将其输入冻结的 T5-XXL 编码器，得到一系列 4096 维的语言嵌入向量。这些嵌入向量经过线性投影，转换为我们的 Transformer 模型（基础模型和超分辨率版）的隐藏层大小。

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 等人，2021b）模型中获得的语义词元。该模型由一个编码器和一个解码器组成，并带有一个量化层，可将输入图像映射为从已学习码本中提取的词元序列。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。编码器包含多个下采样块，用于缩小输入的空间维度；解码器则包含相应数量的上采样块，将潜变量映射回原始图像尺寸。给定尺寸为 $H \times W$ 的图像，编码后的词元尺寸为 ${}^{H}/_{f} \times {}^{W}/_{f}$，下采样率为 f。我们训练了两个 VQGAN 模型：一个的下采样率为 $f = 16$，另一个的下采样率为 $f = 8$。我们使用 $f = 16$ VQGAN 模型对 256×256 像素图像进行编码，以获得基础模型所需的词元，因此词元的空间尺寸为 16×16。我们使用 $f = 8$ VQGAN 模型对 512×512 图像进行编码，以获得超分辨率模型所需的词元，相应词元的空间尺寸为 64×64。如先前研究（Esser 等人，2021b）所述，编码后得到的离散词元能够捕捉图像的高层语义，同时忽略低级噪声。此外，这些词元的离散特性使我们能够在输出端使用交叉熵损失，以预测下一阶段的掩码词元。

2.3. 基础模型

我们的基础模型是一个掩码 Transformer（Vaswani 等人，2017年；Devlin 等人，2018），其输入为投影后的 T5 嵌入和图像令牌。我们不对任何文本嵌入进行掩码，而是随机掩码不同比例的图像令牌（见第2.6节），并将其替换为特殊的 [掩码] 令牌（Chang 等人，2022年）。然后，我们将图像令牌线性映射为符合 Transformer 输入/隐藏层所需尺寸的图像输入嵌入，并加入学习得到的二维位置嵌入。遵循以往的 Transformer 架构（Vaswani 等人，2017年），我们使用若干 Transformer 层，包括自注意力块、交叉注意力块和 MLP 块来提取特征。在输出层，使用一个 MLP 将每个被掩码的图像嵌入转换为一组逻辑值（对应于 VQGAN 码本大小），并以真实令牌标签为目标应用交叉熵损失。在训练过程中，基础模型会在每一步预测所有被掩码的令牌。然而，在推理过程中，掩码预测以迭代方式进行，从而显著提升生成质量。

详见第2.8节。

2.4. 超分辨率模型

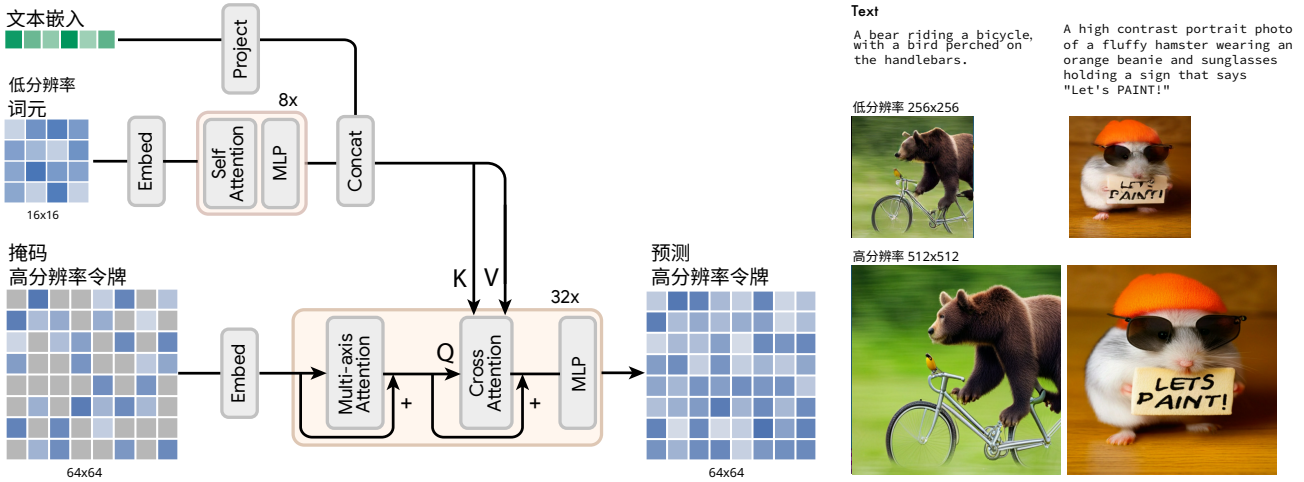


图 *4*。超分辨率模型。左侧展示了超分辨率模型的架构。低分辨率令牌被输入一系列自注意力 Transformer 层；随后，将得到的输出嵌入与从条件文本提示中提取的文本嵌入拼接起来。接着，对这些拼接后的嵌入与掩码高分辨率令牌应用交叉注意力；损失函数通过低分辨率词元和文本词元进行条件化，学习预测这些掩码词元。右侧展示了超分辨率模型带来改进的两个示例。

我们发现，直接预测 512×512 分辨率会导致模型专注于低层次细节，而忽视大规模语义。因此，我们发现采用模型级联很有益：首先由基础模型生成 16×16 潜在映射（对应于 256×256 图像），然后由超分辨率模型将基础潜在映射上采样为64×64 潜在映射（对应于 512×512 图像）。超分辨率模型在基础模型训练完成后进行

训练。
如第 2.2 节所述，我们训练了两个 VQGAN 模型：一个具有 16×16 的潜在分辨率和 256×256 的空间分辨率，另一个具有 64×64 的潜在分辨率和 512×512 的空间分辨率。由于我们的基础模型输出对应于 16×16 潜在映射的词元，因此我们的超分辨率过程会学习将低分辨率潜在映射“转换”为高分辨率潜在映射，随后通过高分辨率 VQGAN 解码，生成最终的高分辨率图像。与基础模型类似，这个潜在映射转换模型也采用文本条件和交叉注意力进行训练，如图 4 所示。

2.5. 解码器微调

为了进一步提升模型生成精细细节的能力，我们通过增加更多残差层和通道来提高 VQGAN 解码器的容量，同时保持编码器容量不变。随后，我们对新的解码器层进行微调，同时冻结 VQGAN 编码器权重、码本和 Transformer（即基础模型和超分辨率模型）。这样，我们无需重新训练其他模型组件（因为视觉词元的“语言”保持不变），就能提升视觉质量。附录中的图 13 展示了这一点：我们可以看到，微调后的解码器能够更清晰地重建店面细节。附录中还介绍了微调解码器的架构细节。

2.6. 可变掩码率

与（Chang 等人，2022年）中的做法相同，我们采用余弦调度，以可变掩码率训练模型：对于每个训练样本，我们从截断的 $r \in [0, 1]$ 反余弦 分布中采样一个掩码率，其概率密度函数为$p(r) = \frac{2}{\pi}(1 - r^2)^{-\frac{1}{2}}$。其期望掩码率为 0.64，并且显著偏向较高的掩码率。较高掩码率带来的偏向使预测问题更难。与针对某种固定词元顺序学习条件分布 $P(x_i|x_{<i})$ 的自回归方法不同，采用可变掩码比例的随机掩码使模型能够针对任意词元子集学习 $P(x_i|x_\Lambda)$。这不仅对并行采样方案至关重要，还支持多种零样本、开箱即用的编辑功能，如图2 和第3.3节所示。

2.7. 无分类器引导

我们采用无分类器引导（CFG）（Ho 和 Salimans，2022年）来提升生成质量和文本图像对齐效果。在训练时，我们随机选取 10% 的样本，移除其中的文本条件（此时注意力退化为图像词元自注意力）。在推理时，对于每个掩码词元，我们计算条件对数几率 ℓ_c 和无条件对数几率 ℓ_u。然后，通过从无条件对数几率出发，沿远离它的方向移动 ℓ_g，得到最终逻辑值 t，其中 引导尺度为：

$$\ell_g = (1 + t)\ell_c - t\ell_u \tag{1}$$

直观来说，CFG 是在多样性和保真度之间进行权衡。与以往方法不同，我们通过在采样过程中线性增大引导尺度 t，来减轻对多样性的损害。这样一来，较早生成的词元可以更自由地采样，接受较弱或不接受引导；而较晚生成的词元则会受到条件提示更大的影响。
我们还利用这一机制实现负向提示（NegPrompt，2022年），方法是将无条件对数几率 ℓ_u替换为以“负向提示”为条件的对数几率。这会促使生成的图像具备与正向提示相关的特征 ℓ_c，并移除与负向提示相关的特征 ℓ_u。

2.8. 推理时的迭代并行解码

我们模型在推理时间效率方面的关键在于采用并行解码，在一次前向传播中预测多个输出词元。并行解码之所以有效，其背后的关键假设是马尔可夫性质：给定其他词元后，许多词元在条件上相互独立。解码基于余弦调度（Chang 等人，2022年），该调度会选出一定比例、置信度最高且要在当前步骤预测的掩码词元。随后，这些词元在剩余步骤中会被设为非掩码状态，并相应减少掩码词元的数量。通过这一流程，我们的基础模型仅需 256 个词元便可用 24 个解码步骤完成推理；我们的超分辨率模型则仅需 4096 个词元和 8 个解码步骤。相比之下，自回归模型需要 256 或 4096 个步骤（例如（Yu 等人，2022年）），而扩散模型则需要数百个步骤（例如（Rombach 等人，2022年；Saharia 等人，2022年））。我们注意到，近期的一些方法，包括渐进式蒸馏（Salimans 和 Ho，2022年）以及



图 *5.* 推理样本。我们展示了基础模型（左）和超分辨率模型（右）在逐步采样过程中掩码令牌的演变。超分辨率模型以低分辨率令牌为条件，因此只需少得多的采样步数即可收敛。

更优秀的常微分方程求解器（Lu 等人，2022年）大幅减少了扩散模型的采样步数，但尚未在大规模文本到图像生成任务中得到广泛验证。我们将与这些更快速方法的比较留待未来工作，同时指出，类似的蒸馏方法也适用于我们的模型。

3. 结果

我们训练了多个不同参数规模的基础 Transformer 模型，参数量从 6 亿到 30 亿不等。这些模型均以 T5-XXL 模型的输出嵌入作为输入；该模型经过预训练后被冻结，包含 46 亿个参数。我们最大的基础模型包含 30 亿个参数，由 48 层 Transformer 组成，具有从文本到图像的交叉注意力，以及图像令牌之间的自注意力。所有基础模型共用同一个图像分词器。我们使用包含 19 个 ResNet 块和大小为8192 的量化码本的卷积神经网络模型进行分词。增大码本大小并未带来性能提升。超分辨率模型由 32 个多轴 Transformer 层（Zhao 等人，2021年）组成，通过将文本和图像嵌入拼接后与高分辨率图像进行交叉注意力，并在高分辨率图像令牌之间进行自注意力。该模型将一个潜在空间中的令牌序列转换到另一个潜在空间：第一个潜在空间是基础模型分词器对应的潜在空间，包含 16×16 个令牌；另一个则是更高分辨率分词器对应的潜在空间，包含 64×64个令牌。完成令牌转换后，使用更高分辨率分词器的解码器将其转换到更高分辨率的图像空间。有关配置的更多细节，请参阅附录。

我们使用由 4.6 亿组文本-图像对组成的 Imagen 数据集进行训练（Saharia 等人，2022年）。训练共进行 100 万步，批量大小为 512，使用 512 核 TPU-v4 芯片（Jouppietal.，2020年）。训练耗时约一周。我们使用 Adafactor 优化器（Shazeer & Stern，2018）以节省内存，从而能够在不进行模型并行化的情况下容纳一个拥有 30 亿参数的模型。我们还避免在训练期间对模型权重进行指数移动平均（EMA），以进一步节省 TPU 内存。为了充分利用 EMA 的优势，我们每训练 5000 步保存一次检查点，随后对检查点中的权重离线执行 EMA，衰减因子设为 0.7。经过平均的权重构成最终基础模型的权重。

3.1. 定性性能

图 6定性展示了 Muse 对具有不同属性的文本提示的处理能力。图 6 左上方的示例展示了模型对基数的基本理解。对于基数大于 1 的物体，Muse 不会简单地重复生成相同的物体像素，而是加入上下文变化，使整体图像更加逼真，例如大象的大小和朝向、酒瓶标签的颜色，以及网球的旋转角度。图 6 右上方展示了模型对多对象构图和相对位置关系的理解。Muse 不会将物体随机放置，而是生成能够保留文本中介词表达的对象关系的图像，例如在……上方与在……下方、左与右等。图 6 左中部展示了模型生成多种风格图像的能力，既包括知名艺术家的特定风格（如伦勃朗），也包括较为普遍的整体风格（如波普艺术和中国水墨画）。图 6 右中部展示了 Muse 呈现词语和短语的能力。文本生成与生成大多数其他物体有着本质区别。模型学习的并非物体名称与其特征之间的映射（例如，“大象”对应“体型大”“灰色”和“吃花生”），因为词语和短语的可能组合近乎无穷，这要求模型采用不同的学习方式。相反，模型必须学习短语、词语和字母之间的层级关系。图 6 左下方展示了 Muse 在渲染时会利用文本提示的全部内容，而非只关注

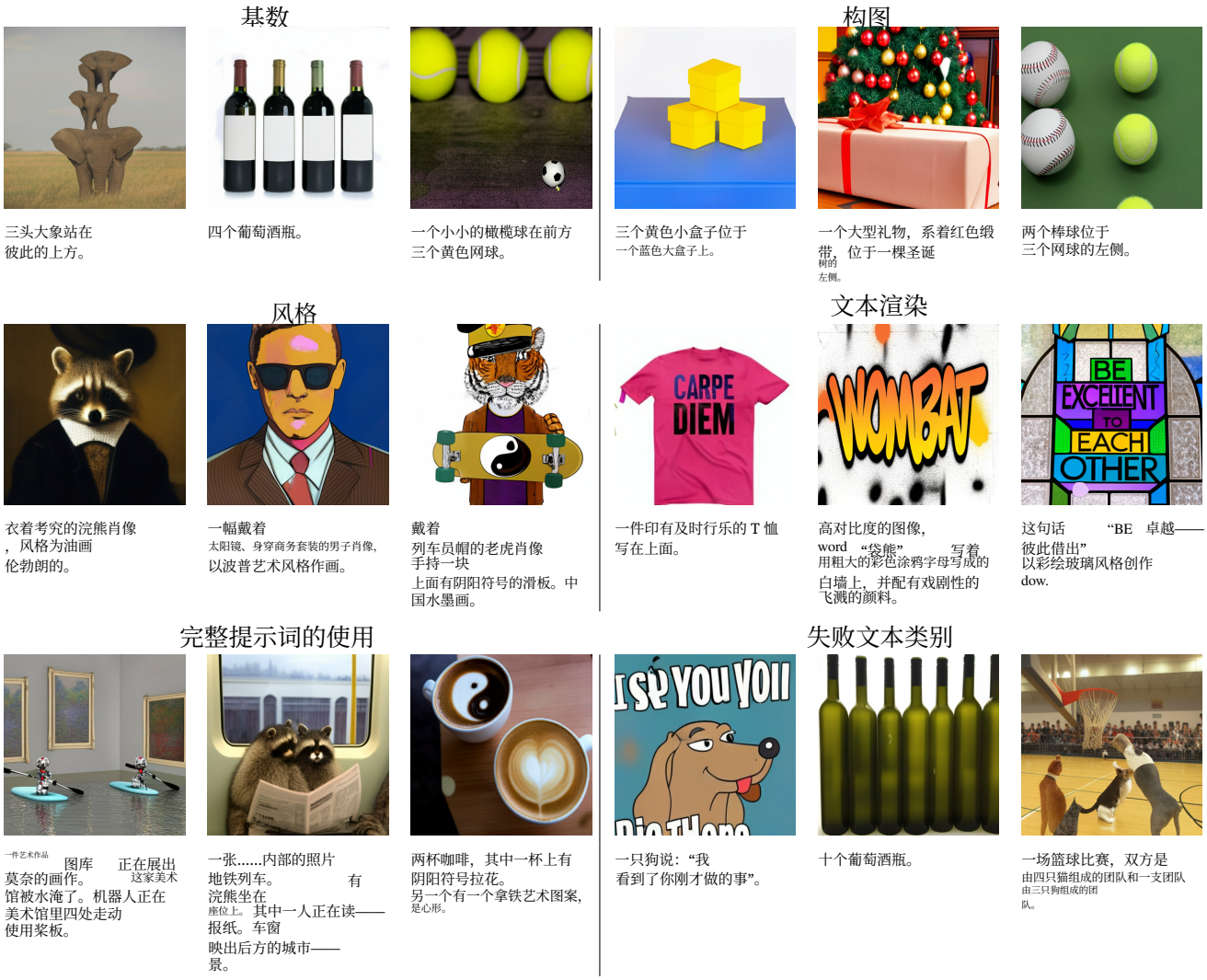


图*6*。展示 Muse 在各种文本属性下文本到图像生成能力的示例。左上：基数；右上：构图；左中：风格；右中：文本渲染；左下：完整提示词的使用情况。对于所有示例，每个提示词生成 16 个实例，并选取 CLIP 分数最高的一个（Radford 等人，2021年）。右下：Muse 在各种文本属性方面生成图像失败的示例，例如长短语的直接渲染、高基数以及多个基数。

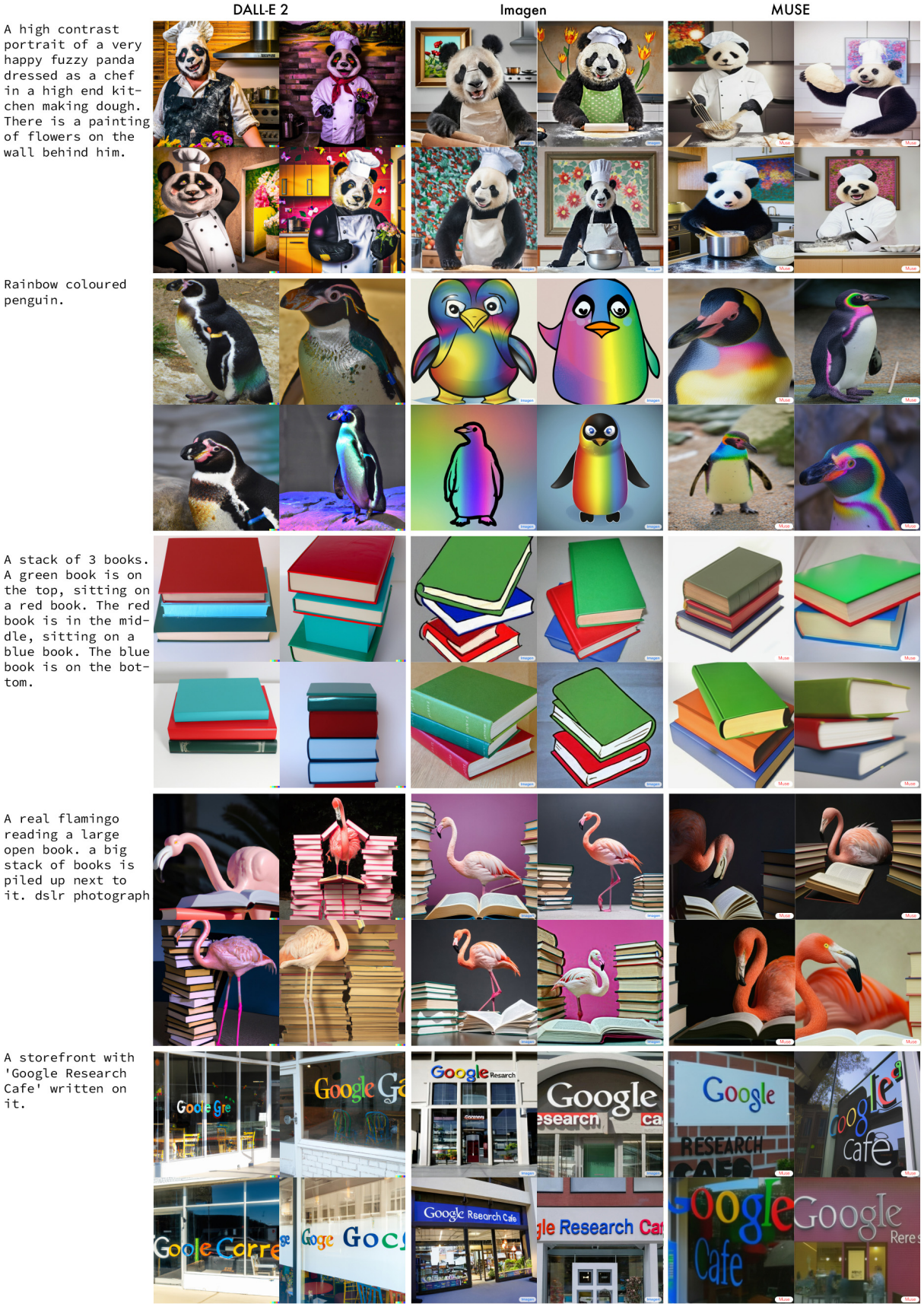


图7。比较相同的提示在 DALL-E2（Ramesh 等人，2022年）（左）、Imagen（Saharia 等人，2022年）（中）和 Muse（右）。

方法	模型类型	参数量	FID	CLIP
VQGAN(Esser 等人, 2021b)	自回归	600M	28.86	0.20
ImageBART(Esser 等人, 2021a)	扩散+自回归	2.8B	22.61	0.23
LDM-4（Rombach 等人，2022年）	扩散	645M	17.01	0.24
RQ-Transformer（Lee 等人，2022a）	自回归	654M	12.33	0.26
草拟与修订（Lee 等人，2022b）	非自回归	654M	9.65	0.26
Muse（基础模型）	非自回归	632M	6.8	0.25
Muse（基础 + 超分辨率版）	非自回归	632M + 268M	6.06	0.26

表*1*。 在 CC3M 上进行定量评估（Sharma 等人，2018）；所有模型均在 CC3M 上训练和评估 .

方法	模型类型	参数量	FID-30K	零样本 FID-30K
AttnGAN（Xu 等人，2017年）	GAN		35.49	-
DM-GAN（Zhu 等人，2019年）	GAN		32.64	-
DF-GAN（Tao 等人，2020年）	GAN		21.42	-
DM-GAN +对比学习（Ye 等人，2021年）	GAN		20.79	-
XMC-GAN（Zhang 等，2021年）	GAN		9.33	-
LAFITE（Zhou 等人，2021年）	GAN		8.12	-
Make-A-Scene（Gafni 等，2022年）	自回归		7.55	-
DALL-E（Ramesh 等人，2021年）	自回归		-	17.89
LAFITE（Zhou 等人，2021年）	GAN		-	26.94
潜在扩散模型（Rombach 等人，2022年）	扩散		-	12.63
GLIDE（Nichol 等人，2021年）	扩散		-	12.24
DALL-E 2（Ramesh 等人，2022年）	扩散		-	10.39
Imagen-3.4B（Saharia 等人，2022年）	扩散		-	7.27
Parti-3B（Yu 等人，2022年）	自回归		-	8.10
Parti-20B（Yu 等人，2022年）	自回归		3.22	7.23
Muse-3B	非自回归		-	7.88

表*2*。 在 MS-COCO（Lin 等人，2014年）上的 FID 和 CLIP 分数（若有）的定量评估，针对 256×256 图像分辨率。Muse 的 CLIP 分数为 0.32，高于 Imagen 报告的 0.27。上表中的其他论文均未报告 CLIP 分数。

仅关注少数几个显著词语。最后，图7 展示了 Muse、Dall-E 2（Ramesh 等人，2022年）和 Imagen（Saharia 等人，2022年针对一些精选提示词的效果对比，结果表明，Muse 与 Imagen 不相上下，并且在许多提示词上的定性表现优于 DALL-E2。

然而，如图6右下角所示，Muse 在生成与某些类型提示词高度匹配的图像方面存在局限。对于要求直接呈现较长多词短语的提示词，Muse 往往会错误地呈现这些短语，常常导致（不必要的）词语重复，或只呈现短语的一部分。此外，表示对象基数较高的提示词往往会生成未能正确体现所需基数的图像（例如，提示词指定10 个葡萄酒瓶，图像却只生成了7个）。总体而言，随着基数增加，Muse 正确呈现物体数量的能力会下降。另一种对 Muse 来说较难处理的提示词，是包含多个基数的提示词（例如“4只猫和一队3只狗”）。对于这类情况，Muse 往往至少会在图像中错误呈现一个数量。

3.2. 定量性能

在表 1 和表 2中，我们展示了我们的方法与其他方法在 CC3M（Sharma 等人，2018）和 COCO（Lin 等人，2014年）数据集上的表现，评估指标为 Fr´echet Inception 距离（FID）（Heusel 等人，2017年），该指标用于衡量样本的质量和多样性；以及 CLIP（Radford 等人，2021年）分数，该分数用于衡量图像/文本对齐。对于 CC3M

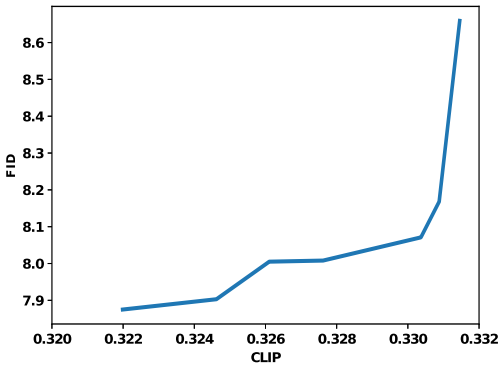
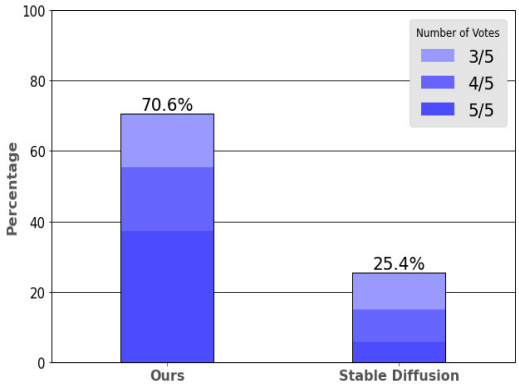


图*8*。CLIP 分数与 FID 分数之间的权衡曲线。我们对固定模型的采样参数进行扫描，然后绘制帕累托前沿。



人工评审共识选择某种模型对齐偏好的提示词百分比。不同颜色表示特定数量的评审共识所作出的选择，而共识的边际百分比（、 和 ）则以数字标示。

结果而言，两个 Muse 模型均在 CC3M 上训练。COCO 的结果为零样本结果，使用与 Imagen 相同数据集上训练的模型（Saharia 等人，2022年）。

我们的 632M 模型在 CC3M 上取得了最先进水平的结果，在 FID 分数方面显著优于现有最先进水平，同时 CLIP 分数也达到最先进水平。我们的 3B 模型取得了 7.88 的 FID 分数，略优于具有相近参数量的 Parti-3B 模型所取得的分数 8。1 。我们的 CLIP 分数为 0.32，高于 Imagen 的 CLIP 分数 0.29（该分数是在 FID 显著更高、达到 20 时取得的）。当 FID 为 7.27 时，Imagen 的 CLIP 分数约为 0.27（见（Saharia 等人，2022年）中的图 4）。

我们的采样算法（第 2.8节）包含多个超参数，例如引导尺度、采样温度，以及是否在采样过程中线性增加引导等。我们对这些参数进行了多轮评估扫描，并找出了帕累托有效的采样参数子集：在这些参数下，无法在不损害 CLIP 分数的情况下改善 FID。这使我们能够研究多样性与图像/文本对齐之间的权衡，如图 8所示。

3.2.1. 人工评估

与以往工作（Yu 等人，2022；Saharia 等人，2022）类似，我们进行了并排评估：向人工评审员展示一个文本提示和两张图像，每张图像均由不同的文本到图像模型根据该提示生成。评审员需要通过回答“哪张图像与图注更匹配？”来评估提示与图像对齐情况。每对图像都经过匿名处理，并随机排列左右顺序。评审员可以选择任一张图像，也可以表示两者没有差别[1]。每组（提示、图像对）由五位独立评审员评估；评审员由 Google 内部众包团队提供，且对 Muse 团队完全匿名。我们使用 PartiPrompts（Yu 等人，2022）作为提供给评审员的提示集。该集合包含 1650 条文本提示，经过精心整理，用于衡量模型在各种类别中的能力。对于两个文本到图像模型，我们将 Muse（30亿参数）与 Stable Diffusion v1.4（Rombach 等人，2022）进行比较；在推理速度方面，Stable Diffusion v1.4 是与 Muse 最为接近的文本到图像模型。对于每条提示，我们生成 16 个图像实例，并选用 CLIP 分数最高的图像（Radford 等人，2021）。Stable Diffusion 图像通过 CompVis Stable Diffusion v1.4 笔记本（CompVis，2022）生成。只有在至少获得 3 位评审员一致同意时，我们才将结果计入相应模型的支持票数。分析结果显示，对于 70.6% 的提示，评审员认为 Muse 的对齐效果优于 Stable Diffusion；对于 25.4% 的提示，评审员认为 Stable Diffusion 的对齐效果优于 Muse；另有 4% 的提示未获得评审员共识。这些结果表明，Muse 的图注匹配能力显著更强（∼2.7倍）。图9展示了评审员分别有 3 票、4 票以及全部 5 票可能投票时的结果细分。对于所有 5 位评审员都认为 Muse 的对齐效果优于 Stable

[1]选择无差异 ma　　当两幅图像都与文本提示不对齐时，这种做法是有意义的，并有助于降低结果中的统计噪声。 s.

扩散是贡献较大的因素。

除了测量对齐程度外，其他研究（Yu 等人，2022年；Saharia 等人，2022年）也测量了图像真实性，通常会通过类似“哪张图像更真实？”的问题来评估。不过，我们注意到，审视此类结果时必须谨慎。尽管问题本意并非如此，但如果某个模型发生了彻底的模式崩溃，无论提示是什么，它都只生成同一张足够真实的图像，那么在回答这个问题时，它几乎总会胜过一个确实会在图像生成时考虑提示的模型。我们认为，这类问题只适用于对齐程度相近的模型之间。由于 Muse 的对齐程度明显优于 Stable Diffusion，我们没有通过人工评审员评估图像真实性。我们认为这一主题仍有待开放研究。

3.2.2. 推理速度

在表3中，我们将 Muse 的推理时间与其他几种热门模型进行了比较。我们在 TPUv4 加速器上对 Parti-3B、Imagen 和 Muse-3B 进行了内部基准测试。对于 Stable Diffusion/LDM，我们采用了已报告的最快基准测试结果（Lambda Labs，2022年），该测试使用 A100 GPU。对于 Stable Diffusion，我们测试的 TPU 实现速度并不比 A100 实现快。我们还报告了 LDM 使用 250 次迭代时的推理时间，这一配置用于达到表 2 中的 FID。尽管 Muse 的参数数量与竞争的扩散模型或自回归模型相当（且参数数量约为 Stable Diffusion/LDM 的 3 倍），其速度仍显著更快。速度 Muse 相较于 Imagen 的优势，源于使用离散令牌并且所需的采样迭代次数更少。Muse 相较于 Parti 的速度优势，源于使用并行解码。Muse 相较于 Stable Diffusion 的速度优势，主要归因于所需的采样迭代次数更少。

模型	分辨率	Time
Imagen	256×256	9.1s
Imagen	1024×1024	13.3秒
潜在扩散模型（50 步）	512×512	3.7s
潜在扩散模型（250 步）	512×512	18.5秒
Parti-3B	256×256	6.4s
Muse-3B	256×256	0.5s
Muse-3B	512×512	1.3s

表3. 多个模型的每批次推理时间。Muse、Imagen 和 Parti 在 TPUv4 硬件上进行了内部基准测试。Stable Diffusion/LDM 的基准测试数据来自（LambdaLabs，2022年），测试使用 A100 GPU。LDM（250 步）的时间是将 50 步的时间乘以 5 得出的；为达到表 2 中的 FID，采用了 250 步。

3.3. 图像编辑

利用模型能够以图像令牌的任意子集为条件这一特性，我们无需额外训练或进行模型微调，即可直接使用该模型处理各种图像编辑应用。

3.3.1. 文本引导的图像内补全／外补全

我们的采样过程（第 2.8 节）可让我们轻松实现文本引导的图像内补全和图像外绘：我们将输入图像转换为一组令牌，遮蔽对应局部区域的令牌，然后以未遮蔽的令牌和文本提示为条件，对掩码令牌进行采样。我们通过多尺度方法集成超分辨率：给定一张尺寸为 512x512 的图像，我们首先将其下采样至 256x256，并将两张图像分别转换为高分辨率和低分辨率词元。然后，遮蔽每组词元中相应的区域。接下来，使用并行采样算法对低分辨率词元进行图像内补全。最后，以这些低分辨率词元为条件，使用相同的采样算法对高分辨率词元进行图像内补全。相关示例见图2 和图10。

3.3.2. 零样本无掩码编辑

我们以零样本方式使用 Muse，对真实输入图像进行无掩码图像编辑。该方法直接作用于（令牌化后的）图像，无需“反演”完整的生成过程；这与近期利用生成模型进行零样本图像编辑的技术不同（Gal 等人，2022b；Patashnik 等人，2021；Kim 等人，2022；Mokady 等人，2022）。

我们首先将输入图像转换为视觉词元。接着，在文本提示的条件下，迭代地掩蔽并重新采样一部分随机选取的词元。我们可以将其视为类似于吉布斯采样过程：固定某些词元，并在这些词元的条件下重新采样其他词元。这样便能使令牌化后的图像进入给定文本提示时图像条件分布的典型集。

我们使用低分辨率基础模型进行编辑，然后对最终输出进行超分辨率处理（以



图*10*。文本引导的图像修复示例。每行第二列显示掩码。这种行为直接源于模型，无需微调。

编辑提示词为条件）。在示例中（图 2、图 11），我们在 100 次迭代中，每次迭代对 8% 的词元进行重采样，引导尺度为 4。我们还对令牌 logits 进行 top-k ($k = 3$) 采样，以防止生成过程偏离输入太多。迭代过程使我们能够控制最终输出。图 12展示了若干中间编辑结果（未经超分辨率处理）；在这个例子中，用户可能更喜欢迭代 50 或 75 次的结果，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 等人，2017年）和生成对抗模型（GAN）已展现出出色的图像生成性能，并且针对卷积式和 Transformer 架构提出了许多变体，例如（Goodfellow 等人，2020年；Esser 等人，2021b；Karras 等人，2019年；Brock 等人，2018；Donahue 和 Simonyan，2019）。直到最近，生成对抗网络一直被视为最先进水平。基于渐进式去噪原理的扩散模型如今能够以同等或更高的保真度合成图像和视频（Ho 等人，2020年；Kingma 等人，2021年；Ho 等人，2022年）。结合多种方法原理的混合方法也展现出出色的性能（Chang 等，2022；Lezama 等人，2022），这表明不同方法之间存在更多可利用的互补性。

4.2. 图像分词器

图像分词器已被证明对多种生成模型都很有用，因为它们能够将大部分计算从输入（像素）空间转移到潜变量空间（Rombach 等人，2022年），或者能够采用更有效的损失函数，例如使用分类而非回归（Chang 等人，2022年；Lezama 等人，2022年；Li 等人，2022年）。目前已发展出多种分词方法，例如离散VAE（Rolfe，2016年）、VQVAE（Van Den Oord 等人，2017年）和 VQGAN（Esser 等人，2021b）；其中，VQGAN 的性能最佳，因为它结合了感知损失与对抗损失，实现了出色的重建效果。ViT-VQGAN（Yu 等人，2021年）将 VQGAN 扩展到了 Transformer 架构。我们使用 VQGAN 而非 ViT-VQGAN，因为我们发现它在我们的模型中表现更好；值得注意的是，分词模型性能更好，并不总意味着文本到图像模型的表现也会更好。

4.3. 大语言模型

我们的工作采用了 T5，这是一种预训练的大语言模型（LLM），经过Raffel 等人于 2020年提出的多种文本到文本任务的训练。大语言模型（包括 T5、BERT（Devlin 等人， 2018）和 GPT（Brown 等人， 2020年；Radford 等人， 2019年））已



图*11*。超分辨率处理后的零样本免掩码图像编辑示例。可以看到，在根据文本提示改变物体的某些特定方面时，图像的姿势和整体结构得以保留。



图*12*。 生成图中某项编辑的中间迭代过程 11 (超分辨率处理前)

已证明其能够学习强大的嵌入，从而支持小样本迁移学习。我们的模型利用了这一能力。所有现代大语言模型都通过词元预测任务进行训练（无论是自回归还是非自回归）。本研究利用了词元预测能力的相关见解，采用 Transformer 来预测视觉词元。

4.4. 文本-图像模型

利用配对文本-图像数据进行表征学习和生成模型训练，正日益展现出强大的学习范式。CLIP（Radford 等人，2021年）和 ALIGN（Jia 等人，2021年）训练模型，使文本和图像嵌入对齐，展现出出色的迁移能力和小样本能力。Imagen（Saharia 等人，2022年）和 Parti（Yu 等人，2022年）使用类似的大规模文本-图像数据集（Schuhmann 等人，2021年；2022年）学习如何根据文本输入预测图像，在 FID 和人工评估中取得出色表现。一项关键技巧是使用无分类器引导（Ho & Salimans，2022年；Dhariwal& Nichol，2021年），以权衡多样性和质量。

4.5. 生成模型图像编辑

生成对抗网络在图像编辑和操控方面的能力已得到广泛研究（综述请参见（Xia 等人，2022年））。研究人员已针对扩散模型开发了多种技术，以实现编辑、个性化和反演至词元空间（Gal 等人，2022a；Meng 等人，2021年；Ruiz 等人，2022年；Kawar 等人，2022年；Brooks 等人，2022年；Hertz 等人，2022年；Mokady 等人，2022年）。DreamBooth（Ruiz 等人，2022年）和 Imagic（Kawar 等人，2022年）都涉及对生成模型进行微调。ImagenEditor（Wang 等人，2022年）将编辑任务定义为文本引导的图像修复，并使用用户指定的掩码。

5. 讨论与社会影响

Muse 模型证实了（Saharia 等人，2022年）的研究发现：冻结的大型预训练语言模型可作为强大的文本到图像生成文本编码器。我们在最初的实验中也尝试从头开始，利用训练数据学习语言模型，但发现其性能显著逊于使用预训练大语言模型，尤其是在处理长提示词和罕见词时。我们还表明，基于 Transformer 架构的非扩散、非自回归模型，其表现可与扩散模型相当，同时推理效率显著更高。我们取得了最先进水平的 CLIP 分数，展现出图像与文本之间出色的对齐效果。我们还通过多种图像编辑应用展示了该方法的灵活性。

我们认识到，生成模型有着多种应用，对人类社会可能产生不同影响。生成模型（Saharia 等人，2022年；Yu 等人，2022年；Rombach 等人，2022年；Midjourney，2022年）具有增强人类创造力的巨大潜力（Hughes 等人，2021年）。然而，众所周知，这些模型也可能被用于传播虚假信息、实施骚扰，以及助长各种社会与文化偏见（Franks & Waldman，2018；Whittaker 等人，2020年；Srinivasan &Uchino，2021年；Steed & Caliskan，2021 年）。鉴于这些重要考量，我们决定目前不发布代码或提供公开演示。

数据集偏见是另一项重要的伦理考量，因为这类模型需要大量数据集，而这些数据集大多是通过自动化方式整理而成。此类数据集可能存在各种问题，例如缺乏同意和受试者知情（Paullada 等人，2021年；Dulhanty，2020年；Scheuerman 等人，2021年）。许多常用数据集往往反映了负面的社会刻板印象和观点（Prabhu & Birhane，2020年）。因此，使用此类数据集进行训练很可能只会放大这些偏见。我们还需要开展大量额外研究，探讨如何减轻这些偏见并生成不含此类偏见的数据集。这是一个非常重要的议题（Buolamwini &Gebru，2018；Hendricks 等人，2018），但超出了本文的讨论范围。

鉴于上述考量，我们不建议在未充分关注各种用例且不了解潜在危害的情况下使用文本生成图像模型。我们尤其不建议使用此类模型生成人物、人类和面部图像。

致谢

感谢 William Chan、Chitwan Saharia 和 Mohammad Norouzi 为我们提供训练数据集、各种评估代码以及慷慨的建议。Jay Yagnik、Rahul Sukthankar、Tom Duerig 和 David Salesin 热情支持了这个项目，对此我们深表感谢。感谢 Victor Gomes 和 Erica Moreira 提供基础设施支持，感谢 Jing

感谢 Yu Koh 和 Jason Baldridge 就数据集、模型和评估进行讨论，并对论文提出反馈；感谢 Mike Krainin 参与模型加速方面的讨论，感谢 JD Velasquez 提供讨论和见解，感谢 Sarah Laszlo、Kathy Meier-Hellstern 和 Rachel Stigler 协助我们完成出版流程，感谢 Andrew Bunner、Jordi Pont-Tuset 和 Shai Noy 协助准备内部演示，感谢 David Fleet、Saurabh Saxena、Jiahui Yu 和 Jason Baldridge 分享 Imagen 和 Parti 的速度指标。

参考文献

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年。Brock, A., Donahue, J., and Simonyan, K. 用于高保真自然图像合成的大规模生成对抗网络训练。 *arXiv*预印本 *arXiv:1809.11096*，2018。Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix：学习遵循图像编辑指令。 *arXiv*预印本 *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., 等。语言模型是少样本学习器。 神经信息处理系统进展，33:1877–1901，2020年。Buolamwini, J. and Gebru, T. 性别阴影：商业性别分类中的交叉性准确率差异。发表于公平性、问责制与透明度会议，第77–91页。机器学习研究论文集，2018。Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit：掩码生成图像Transformer。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第11315–11325页，2022。CompVis。Stable diffusion Colab，2022。网址
h
t
t
p
s://
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：用于语言理解的深度双向Transformer预训练。 *arXiv*预印本 *arXiv:1810.04805*，2018。Dhariwal, P. and Nichol, A. 扩散模型在图像合成方面胜过生成对抗网络。 神经信息处理系统进展，34:8780–8794，2021年。Donahue, J. and Simonyan, K. 大规模对抗式表征学习。 神经信息处理系统进展，32，2019。Dulhanty, C. 计算机视觉数据收集中的问题：偏见、同意与标签分类体系。硕士论文，滑铁卢大学，2020。Esser, P., Rombach, R., Blattmann, A., and Ommer, B. Imagebart：结合多项式扩散实现自回归图像合成的双向上下文。 神经信息处理系统进展，34:3518–3532，2021a。Esser, P., Rombach, R., and Ommer, B. 驯服Transformer以实现高分辨率图像合成。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第12873–12883页，2021b。Franks, M. A. and Waldman, A. E. 性、谎言与录像带：深度伪造与言论自由的迷思。《马里兰法律评论》，78:892，2018。Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene：基于场景并借助人类先验的文本到图像生成，2022。网址
https://arxiv.org/abs/2203.13131。Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. 一张图像胜过千言万语：使用文本反演实现文本到图像生成的个性化。 *arXiv*预印本 *arXiv:2208.01618*，2022a。Gal, R., Patashnik, O., Maron, H., Bermano, A. H., Chechik, G., and Cohen-Or, D. Stylegan-nada：由CLIP引导的图像生成器领域自适应。 《*ACM*图形学汇刊》*(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. 生成对抗网络。 《*ACM*通讯》，63(11):139–144，2020。

Goyal, P., Doll´ar, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. 精确的大批量随机梯度下降：1小时内训练ImageNet。arXiv预印本 arXiv:1706.0267，2017年。He, K., Chen, X., Xie, S., Li, Y., Doll´ar, P., and Girshick, R. 掩码自编码器是可扩展的视觉学习器。发表于计算机视觉与模式识别会议，第16000–16009页，2022年6月。Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. 女性也玩单板滑雪：克服图像描述模型中的偏见。发表于欧洲计算机视觉会议（ECCV）论文集，第771–787页，2018年。Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. 使用交叉注意力控制进行提示到提示图像编辑。arXiv预印本 arXiv:2208.01626，2022年。Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. 采用双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡。神经信息处理系统进展，30，2017年。Ho, J. and Salimans, T. 无分类器扩散引导。arXiv预印本 arXiv:2207.12598，2022年。Ho, J., Jain, A., and Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020年。Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. 视频扩散模型。arXiv预印本 arXiv:2204.03458，2022年。Hughes, R. T., Zhu, L., and Bednarz, T. 生成对抗网络赋能的创意与设计产业人类与人工智能协作应用：当前方法与趋势的系统综述。《人工智能前沿》，4:604234，2021年。Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. 利用噪声文本监督扩展视觉及视觉语言表征学习。发表于国际机器学习会议，第4904–4916页。机器学习研究论文集，2021年。Jouppi, N. P., Yoon, D. H., Kurian, G., Li, S., Patil, N., Laudon, J., Young, C., and Patterson, D. 一种用于训练深度神经网络的领域专用超级计算机。《ACM通讯》，63(7):67–78，2020年。Karras, T., Laine, S., and Aila, T. 用于生成对抗网络的基于风格的生成器架构。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第4401–4410页，2019年。Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic：使用扩散模型进行基于文本的真实图像编辑。arXiv预印本 arXiv:2210.09276，2022年。Kim, G., Kwon, T., and Ye, J. C. Diffusionclip：用于鲁棒图像操控的文本引导扩散模型。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第2426–2435页，2022年。Kingma, D., Salimans, T., Poole, B., and Ho, J. 变分扩散模型。神经信息处理系统进展，34:21696–21707，2021年。Kingma, D. P. and Ba, J. Adam：一种随机优化方法。发表于国际学习表征会议，2015年。Lambda Labs。你只需要一块图形处理器：Stable Diffusion推理基准测试，2022年。网址 https://lambdalabs. com/blog/inference-benchmark-stable-diffusion。Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. 使用残差量化进行自回归图像生成。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第11523–11532页，2022a。Lee, D., Kim, C., Kim, S., Cho, M., and Han, W.-S. 草拟与修订：借助上下文RQ变换器实现有效图像生成。arXiv预印本 arXiv:2206.04452，2022b。Lezama, J., Chang, H., Jiang, L., and Essa, I. 使用Token-Critic改进掩码图像生成。发表于欧洲计算机视觉会议，第70–86页。施普林格，2022年。Li, T., Chang, H., Mishra, S. K., Zhang, H., Katabi, D., and Krishnan, D. MAGE：用于统一表征学习与图像合成的掩码生成编码器。arXiv预印本 arXiv:2211.09117，2022年。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., 和 Zitnick, C. L. Microsoft coco：上下文中的常见物体。载于《欧洲计算机视觉会议》，第 740–755 页。施普林格，2014年。Loshchilov, I. 和 Hutter, F. 随机梯度下降重启：带热重启的随机梯度下降。载于国际学习表征会议，2017年。Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., 和 Zhu, J. Dpm-solver：一种快速常微分方程求解器，可在约 10 步内完成扩散概率模型采样。arXiv 预印本 arXiv:2206.00927，2022年。Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., 和 Ermon, S. Sdedit：使用随机微分方程进行图像合成与编辑。arXiv 预印本 arXiv:2108.01073，2021年。Merullo, J., Castricato, L., Eickhoff, C., 和 Pavlick, E. 从图像空间到文本空间的线性映射。arXiv 预印本 arXiv:2209.15162，2022年。Midjourney。Midjourney，2022年。网址 https://www.midjourney.com。Mokady, R., Hertz, A., Aberman, K., Pritch, Y., 和 Cohen-Or, D. 使用引导式扩散模型编辑真实图像的空文本反演，2022年。网址 https://arxiv.org/abs/2211.09794。NegPrompt。负向提示，2022年。网址 https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt。Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., 和 Chen, M. Glide：借助文本引导扩散模型实现照片级真实感图像生成与编辑。arXiv 预印本 arXiv:2112.10741，2021年。Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., 和 Lischinski, D. Styleclip：以文本驱动对 StyleGAN 图像进行风格操控。载于《IEEE/CVF国际计算机视觉会议论文集》，第 2085–2094 页，2021年。Paullada, A., Raji, I. D., Bender, E. M., Denton, E., 和 Hanna, A. 数据及其（不）满：机器学习研究中数据集开发与使用综述。《Patterns》，2(11):100336，2021年。Prabhu, V. U. 和 Birhane, A. 大型图像数据集：计算机视觉的一场得不偿失的胜利？arXiv 预印本 arXiv:2006.16923，2020年。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 等。语言模型是无监督多任务学习器。《OpenAI 博客》，1(8):9，2019年。Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等。从自然语言监督中学习可迁移视觉模型。载于《国际机器学习会议》，第 8748–8763 页。机器学习研究论文集，2021年。Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., 等。探索使用统一文本到文本转换器进行迁移学习的极限。《机器学习研究杂志》，21(140):1–67，2020年。Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., 和 Sutskever, I. 零样本文本到图像生成，2021年。网址 https://arxiv.org/abs/2102.12092。Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., 和 Chen, M. 基于 CLIP 潜变量的分层文本条件图像生成。arXiv 预印本 arXiv:2204.06125，2022年。Rolfe, J. T. 离散变分自编码器。arXiv 预印本 arXiv:1609.02200，2016年。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 和 Ommer, B. 使用潜在扩散模型进行高分辨率图像合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 10684–10695 页，2022年。Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., 和 Aberman, K. Dreambooth：微调文本到图像扩散模型以实现主体驱动生成。arXiv 预印本 arXiv:2208.12242，2022年。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., 等。借助深度语言理解实现照片级真实感文本到图像扩散模型。arXiv 预印本 arXiv:2205.11487，2022年。

Salimans, T. 和 Ho, J. 用于快速采样扩散模型的渐进式蒸馏。载于国际学习表征会议，2022年。Scheuerman, M. K., Hanna, A., 和 Denton, E. 数据集是否具有政治性？计算机视觉数据集开发中的学科价值观。《美国计算机协会人机交互汇刊》，5(CSCW2):1–37，2021年。Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., 和 Komatsuzaki, A. Laion-400m：由 CLIP 筛选的 4 亿图文对开放数据集。arXiv 预印本 arXiv:2111.02114，2021年。Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 等。Laion-5b：用于训练下一代图文模型的开放式大规模数据集。arXiv 预印本 arXiv:2210.08402，2022年。Sharma, P., Ding, N., Goodman, S., 和 Soricut, R. Conceptual Captions：经过清理并添加上位词的图像替代文本数据集，用于自动图像描述生成。载于《计算语言学协会第 56 届年会论文集（第一卷：长篇论文）》，第 2556–2565 页，2018。Shazeer, N. 和 Stern, M. AdaFactor：具有次线性内存开销的自适应学习率。载于《国际机器学习会议》，第 4596–4604 页。机器学习研究论文集，2018。Srinivasan, R. 和 Uchino, K. 生成艺术中的偏见：从艺术史视角进行因果分析。载于《2021年美国计算机协会公平性、问责制与透明度会议论文集》，第 41–51 页，2021年。Steed, R. 和 Caliskan, A. 通过无监督预训练学习的图像表征包含类人偏见。载于《2021年美国计算机协会公平性、问责制与透明度会议论文集》，第 701–713 页，2021年。Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K., 和 Xu, C. Df-gan：一种简单有效的文本到图像合成基线，2020年。网址 https://arxiv.org/abs/2008.05865。Van Den Oord, A., Vinyals, O., 等。神经离散表征学习。《神经信息处理系统进展》，30，2017年。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., 和 Polosukhin, I. Attention Is All You Need。《神经信息处理系统进展》，30，2017年。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., 和 Chan, W. Imagen Editor 和 EditBench：推进并评估文本引导图像内补，2022年。网址 https://arxiv.org/abs/2212.06909。Whittaker, L., Kietzmann, T. C., Kietzmann, J., 和 Dabirian, A. “我身边全是合成人脸”：人工智能生成媒体的疯狂世界。《IT Professional》，22(5):90–99，2020年。Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., 和 Yang, M.-H. GAN反演：综述。《IEEE模式分析与机器智能汇刊》，2022年。Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., 和 He, X. Attngan：利用注意力生成对抗网络进行细粒度文本到图像生成。CoRR，abs/1711.10485，2017年。网址 http://arxiv.org/abs/ 1711.10485。Ye, H., Yang, X., Takac, M., Sunderraman, R., 和 Ji, S. 使用对比学习改进文本到图像合成，2021年。网址 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., 和 Wu, Y. 通过改进的 VQGAN 进行向量量化图像建模。arXiv 预印本 arXiv:2110.04627，2021年。Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., 等。扩展自回归模型以实现内容丰富的文本到图像生成。arXiv 预印本 arXiv:2206.10789，2022年。Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., 和 Yang, Y. 用于文本到图像生成的跨模态对比学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 833–842 页，2021年。

Zhao, L., Zhang, Z., Chen, T., Metaxas, D. N., and Zhang, H. 用于高分辨率生成对抗网络的改进型Transformer，2021年。URLhttps://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：迈向无需语言的文本到图像生成训练。*CoRR*, abs/2111.13792，2021年。URL https://arxiv.org/abs/2111.13792。Zhu, M., Pan, P., Chen, W., and Yang, Y. DM-GAN：用于文本到图像合成的动态记忆生成对抗网络。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第5802–5810页，2019年。

A. 附录。

A.1. 基础模型配置

我们最大规模模型（30亿参数）的基础模型配置见表4。

配置	值
Transformer 层数	48
Transformer隐藏维度	2048
Transformer多层感知机维度	8192
优化器	AdaFactor(Shazeer& Stern，2018)
基础学习率	1e-4
权重衰减	0.045
优化器动量	$\beta_1=0.9, \beta_2=0.96$
批量大小	512
学习率调度	余弦衰减（Loshchilov&Hutter，2017年）
预热步数	5000
训练步数	1.5M

Table 4. 基础模型的配置和训练超参数 基础模型。

A.2. VQGAN 配置

配置	值
感知损失权重	0.05
对抗损失权重	0.1
码本大小	8192
优化器	Adam（Kingma 和 Ba，2015年）
判别器学习率	1e-4
生成器学习率	1e-4
权重衰减	1e-4
优化器动量	$\beta_1=0.9, \beta_2=0.99$
批量大小	256
学习率调度	余弦衰减（Loshchilov&Hutter，2017年）
预热步数（Goyal 等人，2017年）	10000
训练步数	1M

表 *5.* VQGAN 的配置和训练超参数。

VQGAN 架构：我们的 VQGAN 架构与先前的工作（Esser 等人，2021b）类似。它由若干残差块、下采样（编码器）块和上采样（解码器）块组成。主要区别在于，我们移除了非局部块，使编码器和解码器完全采用卷积式结构，以支持不同的图像尺寸。在基础 VQGAN 模型中，我们在每个分辨率下使用 2 个残差块，基础通道维度为 128。对于微调解码器，我们在每个分辨率下使用 4 个残差块，并将基础通道维度设为 256。



输入图像 VQGAN 重建 微调解码器

图*13*。微调解码器带来的改进示例（第2.5节）。请至少放大至 200%，以观察 VQGAN 重建与使用微调解码器进行重建之间的差异。我们尤其可以看到，门牌号（左下）、店铺招牌（中间）和窗户上的栏杆（右侧）等细节在微调解码器中得到了更好的保留。

A.3. 超分辨率配置

配置	值
低分辨率编码器 Transformer 层数	16
Transformer 层数	32
Transformer隐藏维度	1024
Transformer多层感知机维度	4096
优化器	AdaFactor（Shazeer& Stern，2018）
基础学习率	1e-4
权重衰减	0.045
优化器动量	$\beta_1 = 0.9, \beta_2 = 0.96$
批量大小	512
学习率调度	余弦衰减 (Loshchilov&Hutter，2017年)
预热步数	5000
训练步数	1M

表 *6.* 超分辨率模型的配置和训练超参数。