**Zero-Shot Text-to-Image Generation**

**零样本文本到图像生成**

**摘要**

**Abstract**

Text-to-image generation has traditionally focused on finding better modeling assumptions for training on a fixed dataset. These assumptions might involve complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks supplied during training. We describe a simple approach for this task based on a transformer that autoregressively models the text and image tokens as a single stream of data. With sufficient data and scale, our approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.

文本到图像的生成传统上专注于为固定数据集的训练寻找更好的建模假设。这些假设可能涉及复杂的架构、辅助损失或辅助信息，例如训练期间提供的对象部分标签或分割掩码。我们描述了一种基于转换器的简单方法，该转换器将文本和图像标记自回归建模为单个数据流。凭借足够的数据和规模，我们的方法在以零样本方式评估时与以前的特定领域模型具有竞争力

**1. 介绍**

**1. Introduction**

Modern machine learning approaches to text to image synthesis started with the work of Mansimov et al. (2015), who showed that the DRAW Gregor et al. (2015) generative model, when extended to condition on image captions, could also generate novel visual scenes. Reed et al. (2016b) later demonstrated that using a generative adversarial network (Goodfellow et al., 2014), rather than a recurrent variational auto-encoder, improved image fidelity. Reed et al. (2016b) showed that this system could not only generate objects with recognizable properties, but also could zero-shot generalize to held-out categories.

文本到图像合成的现代机器学习方法始于曼西莫夫等人，他们证明了 DRAW（格雷格等人的生成模型），当扩展到图像说明条件时，也可以生成新颖的视觉场景。里德等人（2016b）后来证明使用生成对抗网络（古德费罗等人2014的一篇论文），而不是循环变分自动编码器，能够提高图像保真度。里德等人（2016b）表明该系统不仅可以生成具有可识别属性的对象，还可以零样本泛化到保留的类别。

Over the next few years, progress continued using a combination of methods. These include improving the generative model architecture with modifications like multi-scale generators (Zhang et al., 2017; 2018), integrating attention and auxiliary losses (Xu et al., 2018), and leveraging additional sources of conditioning information beyond just text (Reed et al., 2016a; Li et al., 2019; Koh et al., 2021).

在接下来的几年中，使用多种方法组合继续取得进展。其中包括通过多尺度生成器等修改来改进生成模型架构（张等人。2017；2018），整合注意力和辅助损失（徐等人。2018），并利用文本以外的其他条件信息来源（芦苇等。，2016a；李等人。，2019；Koh等人。2021)。

Separately, Nguyen et al. (2017) propose an energy-based framework for conditional image generation that obtained a large improvement in sample quality relative to contemporary methods. Their approach can incorporate pretrained discriminative models, and they show that it is capable of performing text-to-image generation when applied to a captioning model pretrained on MS-COCO. More recently, Cho et al. (2020) also propose a method that involves optimizing the input to a pretrained cross-modal masked language model. While significant increases in visual fidelity have occurred as a result of the work since Mansimov et al. (2015), samples can still suffer from severe artifacts such as object distortion, illogical object placement, or unnatural blending of foreground and background elements.

单独地，Nguyen等人。（2017）提出了一个基于能量的条件图像生成框架，与当代方法相比，该框架在样本质量方面有了很大的提高。他们的方法可以结合预训练的判别模型，并且他们表明，当应用于MS-COCO上的预训练解释模型时，它能够执行文本到图像的生成。最近，赵等（2020）还提出了一种方法，该方法涉及优化对预训练的跨模态掩码语言模型的输入。自从这项工作开始以来，视觉保真度显着提高，但曼西莫夫等人（2015）的样本仍然会遭受严重的伪影像，例如对象失真、不合逻辑的对象放置或前景和背景元素的不自然混合。

Recent advances fueled by large-scale generative models suggest a possible route for further improvements. Specifically, when compute, model size, and data are scaled carefully, autoregressive transformers (Vaswani et al., 2017) have achieved impressive results in several domains such as text (Radford et al., 2019), images (Chen et al., 2020), and audio (Dhariwal et al., 2020).

大规模生成模型的最新进展表明了进一步改进的可能途径。具体来说，当计算、模型大小和数据被仔细缩放时，自回归转换器（瓦斯瓦尼等人。2017）在文本（拉德福德等人。2019），图片（陈等人。2020）和音频（达里瓦尔等人。2020）方面取得了令人印象深刻的结果。

By comparison, text-to-image generation has typically been evaluated on relatively small datasets such as MS-COCO and CUB-200 (Welinder et al., 2010). Could dataset size and model size be the limiting factor of current approaches? In this work, we demonstrate that training a 12-billion parameter autoregressive transformer on 250 million image-text pairs collected from the internet results in a flexible, high fidelity generative model of images controllable through natural language.

相比之下，文本到图像的生成通常在相对较小的数据集上进行评估，例如 MS—COCO和CUB—200（韦林德等人。2010）。数据集大小和模型大小会成为当前方法的限制因素吗？在这项工作中，我们展示了在网络上收集的2.5亿图像-文本对上训练的一个120亿参数的自回归变换器，产生了一个灵活、高保真的图像生成模型，可通过自然语言进行控制。

The resulting system achieves high quality image generation on the popular MS-COCO dataset zero-shot, without using any of the training labels. It is preferred over prior work trained on the dataset by human evaluators 90% of the time. We also find that it is able to perform complex tasks such as image-to-image translation at a rudimentary level. This previously required custom approaches (Isola et al., 2017), rather emerging as a capability of a single, large generative model.

由此产生的系统在流行的 MS—COCO 数据集零样本上实现了高质量的图像生成，而无需使用任何训练标签。在90％的时间里，它优于人工评估者在数据集上训练的先前工作。我们还发现它能够在初级水平上执行复杂的任务，例如图像到图像的转换。这以前需要自定义方法（伊索拉等人。2017），而现在是作为单个大型生成模型的能力而出现的。

**2. 方法**

**2. Method**

Our goal is to train a transformer (Vaswani et al., 2017) to autoregressively model the text and image tokens as a single stream of data. However, using pixels directly as image tokens would require an inordinate amount of memory for high-resolution images. Likelihood objectives tend to prioritize modeling short-range dependencies between pixels (Salimans et al., 2017), so much of the modeling capacity would be spent capturing high-frequency details instead of the low-frequency structure that makes objects visually recognizable to us.

我们的目标是训练一个转换器（瓦斯瓦尼等人。2017）将文本和图像标记并自回归建模为单个数据流。但是，直接使用像素作为图像标记将需要大量内存来存储高分辨率图像。似然目标倾向于优先建模像素之间的短程依赖关系（萨利曼斯等人。2017），因此大部分建模能力将用于捕获高频细节，而不是使物体在视觉上被我们识别的低频结构。

We address these issues by using a two-stage training procedure, similar to (Oord et al., 2017; Razavi et al., 2019):

我们通过使用两阶段训练程序来解决这些问题，类似于（奥德等人，2017；拉扎维等人。2019）：

• Stage 1. We train a discrete variational autoencoder (dVAE)1 to compress each 256×256 RGB image into a 32 × 32 grid of image tokens, each element of which can assume 8192 possible values. This reduces the context size of the transformer by a factor of 192 without a large degradation in visual quality (see Fig1ure 1).

·第1阶段。我们训练离散变分自编码器（dVAE）将每个 256x256 RGB 图像压缩成 32x32个图像标记网格，其中每个元素可以假设8192个可能值。这将转换器的上下文大小减少了192倍，而视觉质量没有大幅下降（见图—）。

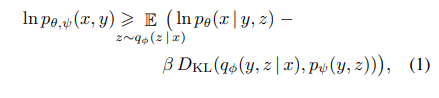
• Stage 2. We concatenate up to 256 BPE-encoded text tokens with the 32 × 32 = 1024 image tokens, and train an autoregressive transformer to model the joint distribution over the text and image tokens.

·第2阶段。我们将多达256个BPE 编码的文本标记与 32x32＝1024个图像标记连接起来，并训练一个自回归转换器来模拟文本和图像标记的联合分布。

The overall procedure can be viewed as maximizing the evidence lower bound (ELB) (Kingma & Welling, 2013; Rezende et al., 2014) on the joint likelihood of the model distribution over images x, captions y, and the tokens z for the encoded RGB image. We model this distribution using the factorization pθ,ψ(x, y, z) = pθ(x | y, z) pψ(y, z), which yields the lower bound

整个过程可以被视为最大化证据下限（ELB）（金马和威灵，2013；Rezende 等人。2014）关于模型分布在图像x、标题 y和编码 RGB 图像的标记z上的联合似然性。我们使用分解

对这个分布进行建模，这会产生下界



where:

• qφ denotes the distribution over the 32 × 32 image tokens generated by the dVAE encoder given the RGB image x;

• pθ denotes the distribution over the RGB images generated by the dVAE decoder given the image tokens;

and

• pψ denotes the joint distribution over the text and image tokens modeled by the transformer.

Note that the bound only holds for β = 1, while in practice we find it helpful to use larger values (Higgins et al., 2016). The following subsections describe both stages in further

detail.

当满足：

·表示给定 RGB 图像x由 dVAE 编码器生成的32 32个图像标记上的分布；

·表示 RGB 图像生成的分布由给定图像令牌的 dVAE 解码器生成；

以及

·表示文本和图像上的联合分布由转换器建模的分类或总结（token）。

注意，该界限仅适用于β＝1，而在实践中，我们发现使用较大的值会有所帮助（希金斯等人。2016）。以下小节更详细地描述了这两个阶段。

**2.1. 第一阶段：学习视觉密码本**

**2.1. Stage One: Learning the Visual Codebook**

In the first stage of training, we maximize the ELB with respect to φ and θ, which corresponds to training a dVAE on the images alone. We set the initial prior pψ to the uniform categorical distribution over the K = 8192 codebook vectors, and qφ to be categorical distributions parameterized by the 8192 logits at the same spatial position in the 32×32 grid output by the encoder.

在训练的第一阶段，我们在φ和方面最大化ELB，这对应于仅在图像上训练 dVAE。我们将初始先验p。设置为K＝8192 密码本向量上的均匀分类分布，并将q。设置为由编码器输出的32x32 网格中相同时空位置的8192个全连接层的输出（logits） 参数化的分类分布。

The ELB now becomes difficult to optimize: as qψ is a discrete distribution, and we cannot use the reparameterization gradient to maximize it. Oord et al. (2017); Razavi et al. (2019) address this using an online cluster assignment procedure coupled with the straight-through estimator (Bengio et al., 2013). We instead use the gumbel-softmax relaxation (Jang et al., 2016; Maddison et al., 2016), replacing the expectation over qφ with one over , where the relaxation becomes tight as the temperature τ → 0. The likelihood for pθ is evaluated using the log-laplace distribution (see Appendix A.3 for a derivation).

ELB 现在变得难以优化：因为是一个离散分布，我们不能使用重新参数化梯度来最大化它。奥德等人。（2017）；拉扎维等人。（2019）使用在线集群分配程序和直通估计器（本吉奥等。2013）。我们改为使用gumbel-softmax 松弛（张等人。2016；麦迪逊等人。 2016），将上的期望替换为上的1，其中松弛度随着温度τ->0变紧。使用对数拉普拉斯分布评估的可能性（参见附录A．3为推导）。

The relaxed ELB is maximized using Adam (Kingma & Ba, 2014) with exponentially weighted iterate averaging. Appendix A.2 gives a complete description of the hyperparameters, but we found the following to be especially important for stable training:

• Specific annealing schedules for the relaxation temperature and step size. We found that annealing τ to 1/16 was sufficient to close the gap between the relaxed validation ELB and the true validation ELB with qφ intsead of  .

• The use of 1 × 1 convolutions at the end of the encoder and the beginning of the decoder. We found that reducing the receptive field size for the convolutions around the relaxation led to it generalizing better to the true ELB.

• Multiplication of the outgoing activations from the encoder and decoder resblocks by a small constant, to ensure stable training at initialization.

松弛化的 ELB 使用 Adam优化器进行最大化（金马＆Ba， 2014），并且具有指数加权选代平均。附录A.2给出了超参数的完整描述，但我们发现以下对于稳定训练尤为重要

·松弛温度和步长的特定退火时间表。我们发现将τ退火到1/16足以缩小松弛验证ELB和真正的验证ELB之间的差距，其中使用，而不是。

·在编码器末端和解码器开头使用1x1卷积。我们发现减小松弛周围卷积的感受域大小可以使其更好地泛化到真正的ELB。

·将来自编码器和解码器resblock的输出激活乘以一个小常数，以确保在初始化时稳定训练。

We also found that increasing the KL weight to β = 6.6 promotes better codebook usage and ultimately leads to a smaller reconstruction error at the end of training.

我们还发现，将散度（Kullback-Leibler (KL)）权重增加到β=6.6促进更好的密码本使用并最终在训练结束后获得更小的重建损失

**2.2. 第一阶段：学习先验**

**2.2. Stage Two: Learning the Prior**

In the second stage, we fix φ and θ, and learn the priordistribution over the text and image tokens by maximizing the ELB with respect to ψ. Here, pψ is represented by a12-billion parameter sparse transformer (Child et al., 2019).

在第二阶段，我们固定φ和，并通过相对于W最大化ELB来学习文本和图像标记的先验分布。这里由一个120亿参数的稀疏变换器表示（查尔德等人。2019）。

Given a text-image pair, we BPE-encode (Sennrich et al., 2015) the lowercased caption using at most 256 tokens with vocabulary size 16,384, and encode the image using 32 × 32 = 1024 tokens with vocabulary size 8192. The image tokens are obtained using argmax sampling from the dVAE encoder logits, without adding any gumbel noise.Finally, the text and image tokens are concatenated and modeled autoregressively as a single stream of data.

给定一个文本-图像对，我们BPE编码（森重奇等人。2015）使用最多256个标记的小写标题词汇量为16，384，并使用32x32=1024个标记，词汇大小为8192。图像标记是使用dVAE编码器输出的argmax采样获得的，没有添加任何gumbel噪声。最后，文本和图像标记被连接起来并自回归地建模为单个数据流。

The transformer is a decoder-only model in which each image token can attend to all text tokens in any one of its 64 self-attention layers. The full architecture is described in Appendix B.1. There are three different kinds of self-attention masks used in the model. The part of the attention masks corresponding to the text-to-text attention is the standard causal mask, and the part for the image-to-image attention uses either a row, column, or convolutional attention mask.

该转换器是一个仅限解码器的模型，其中每个图像标记都可以处理其64个自注意层中的任何一个中的所有文本标记。完整的架构在附录中描述B.1.模型中使用了三种不同的自注意力掩码。与文本到文本注意对应的注意掩码部分是标准因果拖码，图像到图像注意部分使用行、列或卷积注意拖码。

We limit the length of a text caption to 256 tokens, though it is not totally clear what to do for the “padding” positions in between the last text token and the start-of-image token.

One option is to set the logits for these tokens to −∞ in the self-attention operations. Instead, we opt to learn a special padding token separately for each of the 256 text positions. This token is used only when no text token is available. In preliminary experiments on Conceptual Captions (Sharma et al., 2018), we found that this resulted in higher validation loss, but better performance on out-of-distribution captions.

我们将文本标题的长度限制为256个标记，尽管不完全清楚如何处理最后一个文本标记和图像开始标记之间的“填充”位置。一种选择是在自我注意操作中设置这些类别的输出。相反，我们选择为256个文本位置中的每一个分别学习特殊的填充标记。仅当没有文本标记可用时才使用此标记。在概念字幕的初步实验中（夏尔马等。2018），我们发现这会导致更高的验证损失但在分发外字幕上表现更好。

We normalize the cross-entropy losses for the text and image tokens by the total number of each kind in a batch of data. Since we are primarily interested in image modeling, we multiply the cross-entropy loss for the text by 1/8 and the cross-entropy loss for the image by 7/8. The objective is optimized using Adam with exponentially weighted iterate averaging; Appendix B.2 describes the training procedure in more detail. We reserved about 606,000 images for validation, and found no signs of overfitting at convergence.

我们对文本和图像的交叉熵（cross-entropy）损失进行归一化，由一批数据中每种类型的总数来表示。由于我们主要对图像建模感兴趣，我们将文本的交叉损失乘以1/8，图像的交叉损失乘以7/8。目标使用Adam优化器进行优化，采用指数加权选代平均：附录B.2更详细地描述了训练过程。我们保留了大约606，000张图像进行验证，并且在收敛时没有发现过度拟合的迹象

**2.3. 数据采集**

**2.3. Data Collection**

Our preliminary experiments for models up to 1.2 billion parameters were carried out on Conceptual Captions, a dataset of 3.3 million text-image pairs that was developed as an extension to MS-COCO (Lin et al., 2014).

我们在概念字幕上对多达12亿个参数的模型进行了初步实验，概念字幕是一个包含330万个文本图像对的数据集，它是作为MS-COCO的扩展而开发的(林等人。2014）。

To scale up to 12-billion parameters, we created a dataset of a similar scale to JFT-300M (Sun et al., 2017) by collecting 250 million text-images pairs from the internet. This dataset does not include MS-COCO, but does include Conceptual Captions and a filtered subset of YFCC100M (Thomee et al., 2016). As MS-COCO was created from the latter, our training data includes a fraction of the MS-COCO validation images (but none of the captions). We control for this in the quantitative results presented in Section 3 and find that it has no appreciable bearing on the results. We provide further details about the data collection process in Appendix C.

为了扩展到120亿个参数，我们创建了一个与JFT-300M规模相似的数据集（孙小等人。2017）通过从互联网上收集2.5亿对文本图像。该数据集不包括MS-COC0，但包括概念字幕和YFCC100M的过滤子集（托梅等人。2016）。由于MS-COCO是从后者创建的，因此我们的训练数据包括一小部分MS-COCO验证图像（但没有标题）。我们在部分给出的定量结果中对此进行了控制并发现它对结果没有明显影响。我们提供有关数据收集的详细信息在附录C中。

**2.4. 混合精度训练**

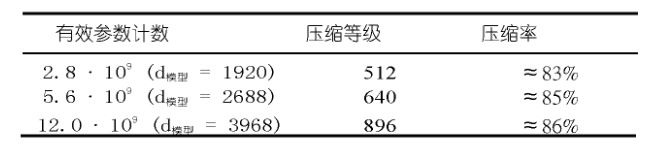
**2.4. Mixed-Precision Training**

To save GPU memory and increase throughput, most parameters, Adam moments, and activations are stored in 16-bit precision. We also use activation checkpointing and recompute the activations within the resblocks during the backward pass. Getting the model to train in 16-bit precision past one billion parameters, without diverging, was the most challenging part of this project.

为了节省GPU内存并提高吞吐量，大多数参数、Adam优化中间参数和激活函数都以16位精度存储。我们还使用激活检查点并在反向传递期间重新计算resblock内的激活让模型以16位精度训练超过10亿个参数，而不发散，是该项目最具挑战性的部分。

We believe the root cause of this instability to be underflow in the 16-bit gradients. Appendix D presents a set of guidelines we developed to avoid underflow when training large-scale generative models. Here, we describe one of these guidelines: per-resblock gradient scaling.

我们认为这种不稳定性的根本原因是16位梯度中的下溢。附录D提出了一组我们制定的指导方针，以避免在训练大规模生成模型时出现下溢。在这里，我们描述了这些准则之一：per-resblock梯度缩放。与之前的工作类似（刘等人。2020），我们发现当我们从较早的resblock移动时，来自前面resblock传给后续resblock的激活梯度的范数单调递减。



Similar to prior work (Liu et al., 2020), we found that the norms of the activation gradients from the resblocks decrease monotonically as we move from the earlier resblocks to the later ones.8 As the model is made deeper and wider, the true exponents of the activation gradients for later resblocks can fall below the minimum exponent of the 16-bit format. Consequently, they get rounded to zero, a phenomenon called underflow. We found that eliminating underflow allowed for stable training to convergence.

随着模型变得更深和更宽，后续resblock的激活梯度的真实指数可能会低于16位格式的最小指数。因此，它们被四舍五入为零，这种现象称为下溢。我们发现消除下溢可以实现稳定的收敛训练。

Standard loss scaling (Micikevicius et al., 2017) is able to avoid underflow when the range spanned by the smallest and largest activation gradients (in absolute value) fits within the exponent range of the 16-bit format. On NVIDIA V100 GPUs, this exponent range is specified by five bits. While this is sufficient for training vanilla language models of the same size, we found the range to be too small for the text-to-image model.

标准损失缩放（Micikevicius等人。2017）当最小和最大激活梯度（绝对值）跨越的范围符合16位格式的指数范围时，能够避免下溢。在NVIDIA V100GPU上，该指数范围由五位指定。虽然这足以训练相同大小的普通语言模型，但我们发现范围对于文本到图像模型来说太小了。

Our fix, which is shown in Figure 4, involves using a separate “gradient scale” for each resblock in the model. This can be seen as a practical alternative to a more general framework for mixed-precision training called Flexpoint (Köster et al., 2017), with the advantage that specialized GPU kernels are not required. We found that Sun et al. (2020) had independently developed similar procedure for training convolutional networks in 4-bit precision.

我们的解决方式 ，如图所示4，涉及为模型中的每个resblock使用单独的梯度比例。这可以看作是一个更通用的混合精度训练框架的替代方案，称为Flexpoint（科斯特等。2017），其优点是不需要专门的GPU内核。我们发现（孙等人。2020）已经独立开发了类似的程序，用于以4位精度训练卷积网络。

**2.5.分布式优化**

**2.5. Distributed Optimizatio**

Our 12-billion parameter model consumes about 24 GB of memory when stored in 16-bit precision, which exceeds the memory of a 16 GB NVIDIA V100 GPU. We address this using parameter sharding (Rajbhandari et al., 2019). As shown in Figure 5, parameter sharding allows us to almost completely hide the latency of the intra-machine communication by overlapping it with compute-intensive operations.

我们的120亿参数模型以16位精度存储时会消耗大约24GB的内存，这超过了16GB NVIDIA V100GPU的内存。我们使用参数分片来解决这个问题（拉态班达里等人。2019）。如图5，参数分片允许我们通过与计算密集型操作重叠来几乎完全隐藏机器内的通信。

On the cluster used to train the model, the bandwidth between machines is much lower than the bandwidth among GPUs on the same machine. This makes the cost of the operation used to average the gradient among the machines (all-reduce) the main bottleneck during training. We were able to drastically reduce this cost by compressing the gradients using PowerSGD (Vogels et al., 2019).

在用于训练模型的集群上，机器之间的带宽远低于同一机器上GPU之间的带宽。这使得用于平均机器之间梯度的操作（all-reduce）成本成为训练期间的主要瓶颈。我们曾经能够通过使用PowerSGD压缩梯度来大降低成本（沃格尔斯等人。，2019）。

In our implementation, each GPU in a machine computes the low-rank factors for its parameter shard gradients independently of its neighboring GPUs.9 Once the low-rank factors are computed, each machine sets its error buffer to the residual between the uncompressed gradient averaged over its eight GPUs (obtained from reduce-scatter), and the decompressed gradient obtained from the low-rank factors.

在我们的实现中，机器中的每个GPU独立于其相邻GPU计算其参数分片梯度的低秩因子。一旦计算了低秩因子，每台机器将其误差缓冲区设置为在其八个GPU上平均的未压缩梯度（从减少散射获得）与从低秩因子获得的解压缩梯度之间的残差。

PowerSGD replaces the large communication operation for an uncompressed parameter gradient with two, much smaller communication operations for its low-rank factors. For a given compression rank r and transformer activation size dmodel, the compression rate is given by 1 − 5r/(8dmodel) (see Appendix E.1). Table 1 shows that we can achieve a compression rate of about 85%, independent of model size.

PowerSGD用两个针对其低秩因子的小得多的通信操作代替了针对未压缩参教梯度的大型通信操作。对给定的压缩等级r和转换器激活大小d模型，压缩率由15r/（8d型）给出（见附录E.1）。表1表明我们可以实现约85%的压缩率，与模型大小无关。

In Appendix E.2, we describe various details that were necessary to get PowerSGD to perform well at scale. These include:

• Saving memory by accumulating the gradient into the error buffers during backpropagation, rather than allocating separate buffers.

• Minimizing instances in which we zero out the error buffers (e.g., due to nonfinite values encountered during mixed-precision backpropagation, or when resuming training from a checkpoint).

• Improving numerical stability by using Householder orthogonalization instead of Gram-Schmidt, together with the addition of a small multiple of the identity matrix to the input.

• Avoiding underflow by using a custom 16-bit floating point format for the error buffers, their low-rank factors, and the all-reduce communication operations involving them.

在附录中E.2，我们描述了使PowerSGD大规模运行所必需的各种细节。这些包括：

·通过在反向传播期间将梯度累积到错误缓冲区中来节省内存，血不是分配单独的缓冲区。

·最小化我们将错误缓冲区归零的情况（例如，由于在混合精度反向传播期间遇到的非有限值，或者在从检查点恢复训练时）。

·通过使用Householder正交化而不是 Gram-Schmidt以及在输入中添加少量单位矩阵来提高数值稳定性。

·通过对错误缓冲区、它们的低秩因子以及涉及它们的all-reduce通信操作使用自定义的16位浮点格式来避免下溢。

We also found the warm-start procedure for the Q matrix described in Vogels et al. (2019) to be unnecessary: we were able to get equivalent results by fixing Q to a random gaussian matrix at the start of training, and never updating it.

我们还发现了Q矩阵的热启动过程沃格尔斯等人（2019）是不必要的：我们能够通过在训练开始时将固定为随机高斯矩阵来获得等效的结果，并且永远不会对其进行更新。

**2.6.样本生成**

**2.6. Sample Generation**

Similar to Razavi et al. (2019), we rerank the samples drawn from the transformer using a pretrained contrastive model (Radford et al., 2021). Given a caption and a candidate image, the contrastive model assigns a score based on how well the image matches the caption. Figure 6 shows the effect of increasing the number of samples N from which we select the top k images. This process can be seen as a kind of language-guided search (Andreas et al., 2017), and is also similar to the auxiliary text-image matching loss proposed by Xu et al. (2018). Unless otherwise stated, all samples used for both qualitative and quantitative results are obtained without temperature reduction (i.e., using t = 1) (except for Figure 2) and use reranking with N = 512.

如同拉扎维等人。（2019），我们使用预训练的对比模型（拉德福德等人。2021）。给定标题和候选图像，对比模型根据图片与标题的匹配程度分配分数。图片6显示了增加样本数量N的效果，我们从中选择前k个图像。这个过程可以看作是一种语言引导的搜索（安德烈亚斯等人。2017），也类似于（徐等人。2018）。除非另有说明，所有用于定性和定量结果的样品都是在没有降低温度的情况下获得的（即使用τ=1）（除了图2）并使用N=512的重新排序。

**3.实验**

**3. Experiments**

**3.1.定量结果**

**3.1. Quantitative Results**

We evaluate our model zero-shot by comparing it to three prior approaches: AttnGAN (Xu et al., 2018), DMGAN (Zhu et al., 2019), and DF-GAN (Tao et al., 2020), the last of which reports the best Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (Heusel et al., 2017) on MS-COCO. Figure 3 qualitatively compares samples from our model to those from prior work.

我们通过将其与三种先前方法进行比较来评估我们的模型零样本：AttnGAN（徐等人。2018），德曼甘（朱等人。2019）和DF-GAN（陶等人。2020），最后一个的精度分数最高（萨利曼等。2016）和Frechet起始距离（赫塞尔等入。，2017）在MS-OCO。数字3定性地将我们模型中的样本与先前工作中的样本进行比较。

We also conduct a human evaluation similar to the one used in Koh et al. (2021) to compare our approach to DF-GAN, the results of which are shown in Figure 7. Given a caption, the sample from our model receives the majority vote for better matching the caption 93% of the time. It also receives the majority vote for being more realistic 90% of the time.

我们还进行了类似于（Koh等人。2021）的仍评价来比较我们与DF-GAN的方法，其结果如图7。给定一个标题，我们模中的样本在93%的情况下获得了多数票，认为该标更好地匹配。它还获得了90%的时间更现实的多数票。

Figure 9(a) shows that our model also obtains an FID score on MS-COCO within 2 points of the best prior approach, despite having never been trained on the captions. Our training data incorporates a filtered subset of YFCC100M, and we found that it includes about 21% of the images in the MS-COCO validation set from a de-duplication procedure described in the next section. To isolate this effect, we compute the FID statistics for the validation set both with these images (solid lines) and without them (dashed lines), finding no significant change in the results.

图9（a）表明我们的模型在MS-COCO上的FID分数也在最佳先前方法的2分之内，尽管从未接受过字幕训练。我们的训练数据包含YFCC100M的过滤子集，我们发现它包含MS-COCO验证集中大约21%的图像这些图像来自下一节中描述的重复数据删除过程。为了隔离这种影响，我们计算了验证集的FID统计数据，包括这些图像（实线）和没有这些图像（虚线），发现结果没有显着变化。

Training the transformer on the tokens from the dVAE encoder allows us to allocate its modeling capacity to the low-frequency information that makes images visually recognizable to us. However, it also disadvantages the model, since the heavy compression renders it unable to produce high-frequency details. To test the effect of this on the quantitative evaluations, we compute the FID and IS in Figure 9(a) after applying a Gaussian filter with varying radius to both the validation images and samples from the models. Our approach achieves the best FID by a margin of about 6 points with a slight blur of radius 1. The gap between our approach and others tends to widen as the blur radius is increased. We also obtain the highest IS when the blur radius is greater than or equal to two.

在来自dVAE编码器的标记上训练转换器，使我们能够将其建模能力分配给低频信息，从而使我们可以在视觉上识别图像。然而，它也对模型不利，因为重度压缩使其无法产生高频细节。为了测试这对定量评估的影响，我们计算了图9（a）中的FID和IS在对验证图像和模型样本应用具有不同半径的高斯滤波器之后。我们的方法以大约6个点的边距实现了最佳FID，半径为1的轻微模糊。随着模湖半径的增加，我们的方法与其他方法之间的差距也随之扩大。当模半径大于或等于2时，我们也获得了最高的IS。

Our model fares significantly worse on the CUB dataset, for which there is a nearly 40-point gap in FID between our model and the leading prior approach (Figure 9(b)). We found an 12% overlap rate for this dataset, and again observed no significant difference in the results after removing these images. We speculate that our zero-shot approach is less likely to compare favorably on specialized distributions such as CUB. We believe that fine-tuning is a promising direction for improvement, and leave this investigation to future work. Samples from our model for captions in this dataset are shown in Figure 8.

我们的模型在CUB数据集上的表现要差得多，我们的模型与先前提出的领先方法之间的FID存在近40个点的差距（图9（b））。我们发现该数据集的重复率为12%并且在移除这些图像后再次观察到结果没有显着差异。我们推测，我们的零样本方法不太可能在CUB等专业发行版上获得有利的比较。我们认为微调是一个有希望的改进方向，并将这项调查留给未来的工作。该数据集中我们的模型中的字幕样本如图8所示。

Finally, Figure 9(c) shows clear improvements in FID and IS for MS-COCO as the sample size used for reranking with the contrastive model is increased. This trend continues up to a sample size of 32, after which we observe diminishing returns.

最后，如图9（c）显示MS-COCO的FID和IS明显改善，因为用于使用对比模型重新排序的样本量增加了。这种超势一直持续到32个样本量，之后我们观察到返回值逐渐降低。

**3.2. .数据重叠分析**

**3.2. Data Overlap Analysis**

We used the deduplication procedure described in Radford et al. (2021) to determine which images to remove. For each validation image, we find the closest image in the training data using a contrastive model specifically trained for this task. We then sort the images in descending order by closeness to their nearest matches in the training data. After inspecting the results by hand, we determine the images to remove by manually selecting a conservative threshold designed to minimize the false negative rate.

我们使用了（拉德福等。2021）描述的重复数据删除过程来确定要删除的图像。对于每个验证图像，我门使用专门为此任务训练的对比模型在训练数据中找到最接近的图像。然后，我们按照与训练数据中最接近的匹配项的接近程度对图像进行降序排序。在手动检查结果后，我们通过手动选择一个保守值来确定要删除的图像，该值旨在最小化假阴性率。

**3.3. 定性发现**

**3.3. Qualitative Findings**

We found that our model has the ability to generalize in ways that we did not originally anticipate. When given the caption “a tapir made of accordion...” (Figure 2a), the model appears to draw a tapir with an accordion for a body, or an accordion whose keyboard or bass are in the shape of a tapir’s trunk or legs. This suggests that it has developed a rudimentary ability to compose unusual concepts at high levels of abstraction.

我们发现我们的模型能够以我们最初没有预料到的方式进行泛化。当给出标题“手风琴制成的貘”时（图2a），该模型似乎在画一条带有手风琴的貘，或者是一只手风琴，其琴键或底座是貘的躯干或腿的形状。这表明它已经发展出一种基本的能力，可以在高度抽象的情况下组合不寻常的概念。

Our model also appears to be capable of combinatorial generalization, such as when rendering text (Figure 2b) or when probed on sentences like “an illustration of a baby hedgehog in a christmas sweater walking a dog” (Figure 2c). Prompts like the latter require the model to perform variable binding (Smolensky, 1990; Greff et al., 2020) – it is the hedgehog that is in the christmas sweater, not the dog. We note, however, that the model performs inconsistently on the task, sometimes drawing both animals with christmas sweaters, or drawing a hedgehog walking a smaller hedgehog.

我们的模型也表现出能够进行组合泛化，例如在渲染文本时（图2b）或当研究诸如“穿着圣诞毛衣的小刺猜狗的插图”这样的句子时（图2c）。像后者一样需要模型执行变量绑定（斯摩棱斯基，1990：格雷夫等人。2020)圣诞毛衣里穿的是刺猬，而不是狗。然而，我们注意到，该模型在任务上的表现不一致，有时会画出两只动物都穿着圣诞毛衣，或者画出一只刺猬走着一只较小的刺猬。

To a limited degree of reliability, we also find our model to be capable of zero-shot image-to-image translation controllable by natural language (Figure 2d). When the model is given the caption “the exact same cat on the top as a sketch at the bottom” and the top 15 × 32 part of the image token grid for a photo of a cat, it is able to draw a sketch of a similar looking cat on the bottom.

在有限的可靠性下，我们还发现我们的模型能够通过自然语言控制零镜头图像到图像的转换（图2d)。当模型被赋予标题“与顶部的猫完全相同的猫从底部看的草图”以及图像标记网格的顶部15x32部分用于猫的照片时，它能够绘制类似的草图从底部看猫。

This works with several other kinds of transformations, including image operations (e.g., changing the color of the image, converting it to grayscale, or flipping it upside-down) and style transfer (e.g., drawing the cat on a greeting card, a postage stamp, or a cell phone case). Some transformations, such as those that involve only changing the color of the animal, suggest that the model is capable of performing a rudimentary kind of object segmentation. We provide additional examples of zero-shot image-to-image translation in Section G.

这适用于其他几种转换，包括图像操作（例如，更改图像的颜色、将其转换为灰度或将其倒置)和样式转换（例如，画猫在贺卡上、在邮票上或在手机壳上）。一些转换，例如只涉及改变动物颜色的转换，表明该模型能够执行一种基本的对象分割。我们在章节中提供了零镜头图像到图像转换的其他示例，见G部分

**4. 结论**

**4. Conclusion**

We investigate a simple approach for text-to-image generation based on an autoregressive transformer, when it is executed at scale. We find that scale can lead to improved generalization, both in terms of zero-shot performance relative to previous domain-specific approaches, and in terms of the range of capabilities that emerge from a single generative model. Our findings suggest that improving generalization as a function of scale may be a useful driver for progress on this task.

基于大规模执行，我们研发了一种基于自回归转换器的简单文本到图像生成方法。我们发现规模可以提高模型的泛化能力，无论是相对于以前的特定领域方法的零样本性能，还是单个生成模型的能力范围。我们的研究结果表明，根据规模改进泛化可能是推动这项任务取得进展的有用驱动力。