CogView: Mastering Text-to-Image Generation via Transformers

CogView: 掌握通过变形器生成文本到图像的方法

摘要

长期以来，一般领域的文本-图像生成一直是一个开放的问题，它需要一个强大的生成模型和跨模式的理解。我们提出了CogView，一个带有VQ-VAE标记器的40亿个参数的转化器来推进这个问题。我们还展示了各种下游任务的微调策略，如风格学习、超级分辨率、文本-图像排名和时装设计，以及稳定预训练的方法，如消除NaN损失。CogView在模糊的MS COCO数据集上实现了最先进的FID，超过了以前基于GAN的模型和最近的类似工作DALL-E。

1介绍

由于对比性自监督预训练已经彻底改变了计算机视觉（CV）[24, 21, 8, 32]，为图像带来高级语义的视觉语言预训练正在成为视觉理解的下一个前沿领域[38, 30, 39]。在各种预文本任务中，文本到图像的生成期望模型能够（1）从像素中分离出形状、颜色、手势和其他特征；（2）理解输入文本；（3）将物体和特征与相应的词及其同义词对齐；（4）学习复杂的分布以生成不同物体和特征的重叠和复合，这和绘画一样，超出了基本的视觉功能（与眼睛和大脑中的V1-V4有关[22]），需要更高级别的认知能力（更与大脑的角回有关[3]）。

教机器进行文字到图像生成的尝试可以追溯到深度生成模型的早期，当时Mansimov等人[35]在DRAW[20]中加入了文字信息。然后，生成对抗网[19]（GANs）开始主导这项任务。Reed等人[42]将文本嵌入作为额外的输入送入生成器和判别器。StackGAN[54]将生成过程分解为一个草图精简过程。AttnGAN[51]使用对单词的关注来关注相应的子区域。ObjectGAN[29]按照文本→方框→布局→图像的过程生成图像。DM-GAN[55]和DF-GAN[45]引入了新的架构，如二元记忆或深度融合块，以实现更好的图像细化。尽管这些基于GAN的模型可以在简单和特定领域的数据集中进行合理的合成，例如Caltech-UCSD Birds 200 (CUB)，但在复杂和一般领域的场景中，例如MS COCO [31]，其结果却远远不能令人满意。

最近几年，自动回归生成模型兴起。生成预训练（GPT）模型[37，4]利用变形金刚[48]在大规模语料库中学习语言模型，极大地促进了自然语言生成和少许语言理解的性能[33]。自动回归模型在CV中并非新生事物。PixelCNN、PixelRNN[47]和Image Transformer[36]用不同的网络骨干对图像的子像素（一个像素中的颜色通道）的概率密度函数进行因子化处理，显示出良好的效果。然而，真实的图像通常由数百万个子像素组成，表明大型模型的计算量是难以承受的。即使是最大的像素级自动回归模型ImageGPT[7]，也是在ImageNet上预训练的，最大分辨率只有96×96。

Vector Quantized Variational AutoEncoders (VQ-VAE) [46] 的框架缓解了这个问题。VQ-VAE训练一个编码器将图像压缩到一个低维的离散潜伏空间，并训练一个解码器从第一阶段的隐藏变量中恢复图像。然后在第二阶段，一个自动回归模型（如PixelCNN[47]）学习适应隐藏变量的先验。这种离散压缩比直接降采样损失的保真度要小，同时还能保持像素的空间相关性。因此，VQ-VAE振兴了CV中的自动回归模型[41]。在此框架下，Esser等人[15]使用Transformer来拟合先验，并进一步将解码器训练从L2损失转换为GAN损失，极大地提高了特定领域无条件生成的性能。

CogView的想法是自然而然的：对文本和图像（来自VQ-VAE）标记的大规模生成性联合预训练。我们收集了3000万个高质量的（中文）文本-图像对，并对一个具有40亿个参数的转化器进行预训练。然而，由于数据的异质性，大规模的文本-图像生成性预训练可能非常不稳定。我们系统地分析了原因，并通过提出的 "精确瓶颈放松 "和 "夹心层规范 "解决了这个问题。因此，CogView大大推进了文本到图像生成的质量。

最近的一项工作DALL-E[39]独立提出了同样的想法，并且比CogView更早发布。与DALL-E相比，CogView在以下四个方面有所进步。

- 根据Fréchet Inception Distance（FID）[25]，CogView在模糊的MS COCO上以较大的优势胜过DALL-E和以前基于GAN的方法，并且是第一个开源的大型文本-图像转换器。

- 除了零点生成，我们还进一步研究了对预训练的CogView进行微调的潜力。CogView可以适用于不同的下游任务，如风格学习（特定领域的文本到图像）、超分辨率（图像到图像）、图像标题（图像到文本），甚至文本-图像重新排序。

- 经过微调的CogView能够为后期选择进行自我排序，并在DALL-E中摆脱了一个额外的CLIP模型[38]。它还提供了一个新的指标Caption Loss来衡量文本-图像生成的质量和准确性，其粒度比FID和Inception Score（IS）更细[43]。

- 我们提出了PB-relaxation和Sandwich-LN来稳定复杂数据集上大型变形器的训练。这些技术非常简单，可以消除转发中的溢出（特点是NaN损失），并使CogView能够以几乎FP16（O22）进行训练。它们也可以被推广到其他变形器的训练中。

2方法

2.1理论

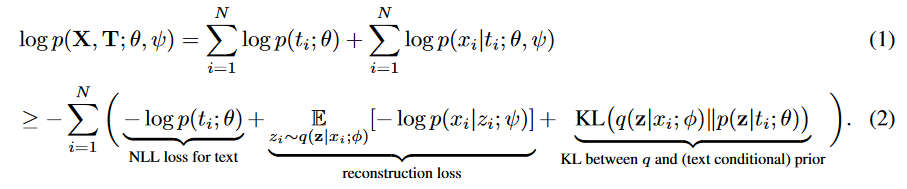
在本节中，我们将从VAE3[26]中推导出CogView的理论。CogView优化了图像和文本的联合似然的证据下限（ELBO）。假设数据集（X，T）={xi，ti}N i=1由图像变量x和其描述的文本变量t的N个i.i.d.样本组成，我们假设图像x可以由涉及潜在变量z的随机过程生成：(1)ti首先由先验p(t; θ)生成。

(2)然后从条件分布p(z|t = ti; θ)生成zi。

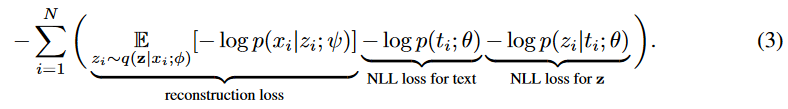
(3)xi最后由p(x|z = zi; ψ)生成。

由先验生成文本t，文本生成潜在变量z，由潜在变量z生成图像x

在下面的部分，我们将使用p(xi)这样的速记形式来指代p(x = xi)。让q(z|xi; φ)为变异分布，它是VAE的编码器φ的输出。对数似然和证据下限（ELBO）可写为



VQ-VAE的框架与传统VAE的不同之处主要在于KL项。传统的VAE固定先验p(z|ti; θ)，通常为N(0, I)，并学习编码器φ。然而，它导致了后验塌陷[23]，即q(z|xi; φ)有时会向先验塌陷。VQ-VAE转向固定φ，用另一个以θ为参数的模型来拟合先验p(z|ti; θ)。这种技术消除了后验塌陷，因为编码器φ现在只为优化重建损失而更新。作为交换，近似的后验q(z|xi; φ)对于不同的xi可能是非常不同的，所以我们需要一个非常强大的p(z|ti; θ)的模型来最小化KL项。目前，最强大的生成模型，Transformer（GPT），应对的是离散编码本上的标记序列。为了使用它，我们让z∈{0, ..., |V | - 1}h×w，其中|V |是编码本的大小，h×w是z的维数。序列zi可以从q(z|xi; φ)中采样，或者直接zi = argmaxz q(z|xi; φ)。为了简单起见，我们选择后者，这样q(z|xi; φ)就成为zi上的一个单点分布。方程（2）可以改写为



然后，学习过程分为两个阶段：

(1) 编码器φ和解码器ψ学习以最小化重建损失。

(2) 单个GPT通过串联文本ti和zi作为输入序列来优化两个负对数似然（NLL）损失。

因此，第一阶段退化为一个纯粹的离散自动编码器，作为一个图像标记器，将图像转换为标记序列；第二阶段的GPT承担了大部分建模任务。

图3说明了CogView的框架。

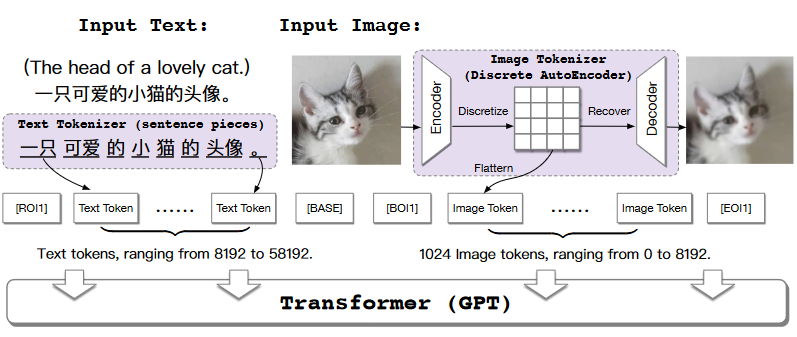
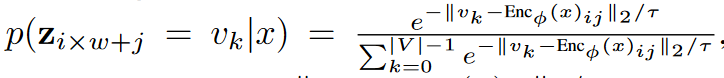


图3：CogView的框架。[ROI1]、[BASE1]等，是分隔符号。

2.2 标记化  
在这一节中，我们将介绍CogView中tokenizer的细节，以及关于image tokenizer（VQVAE阶段1）的不同训练策略的比较。文本的标记化已经得到了很好的研究，例如BPE[16]和SentencePiece[28]。在CogView中，我们在一个大型中文语料库上运行SentencePiece，提取了50,000个文本标记。图像标记器是一个离散的自动编码器，它类似于VQ-VAE[46]或d-VAE[39]的第一阶段。更具体地说，编码器φ将形状为H×W×3的图像x映射为形状为h×w×d的Encφ(x)，然后每个d维向量被量化为可学习编码本{v0, ..., v|V |-1}中的一个附近嵌入，∀vk∈Rd。量化后的结果可以用嵌入的h×w指数表示，然后我们得到潜伏变量z∈{0, ..., |V | - 1}h×w。解码器ψ将量化的向量映射回（模糊的）图像以重建输入。在我们的4B参数CogView中，|V | = 8192，d = 256，H = W = 256，h = w = 32。由于离散选择的存在，图像标记器的训练是不难的。这里我们介绍四种方法来训练图像标记器。  
- 最近邻居映射，直通式估计器[2]，这是由最初的VQVAE提出的。这种方法[39]的一个共同关注点是，当编码本很大且没有仔细初始化时，由于维度的诅咒，只有少数的嵌入会被使用。我们在实验中没有观察到这种现象。  
- Gumbel采样，直通式估计器。如果我们按照原来的VAE来重新参数化基于向量间距离的潜变量z的分类分布，即



一个无偏抽样策略是zi×w+j = argmaxk gk - ‖vk - Encφ(x)ij‖2/τ，gk∼Gumbel(0, 1)，其中温度τ逐渐降低到0。我们可以进一步使用可微调的softmax来近似argmax的一热分布。DALL-E采用这种方法与其他许多技巧来稳定训练。

- 最近的邻居映射，移动平均，在训练过程中，编码本中的每个嵌入被定期更新为最近映射到它的向量的平均值[46]。

- 近邻映射，固定编码本，编码本在初始化后是固定的。

**比较** 为了比较这些方法，我们在相同的数据集和随机种子上训练四个具有相同架构的图像标记器，并在图2中展示损失曲线。

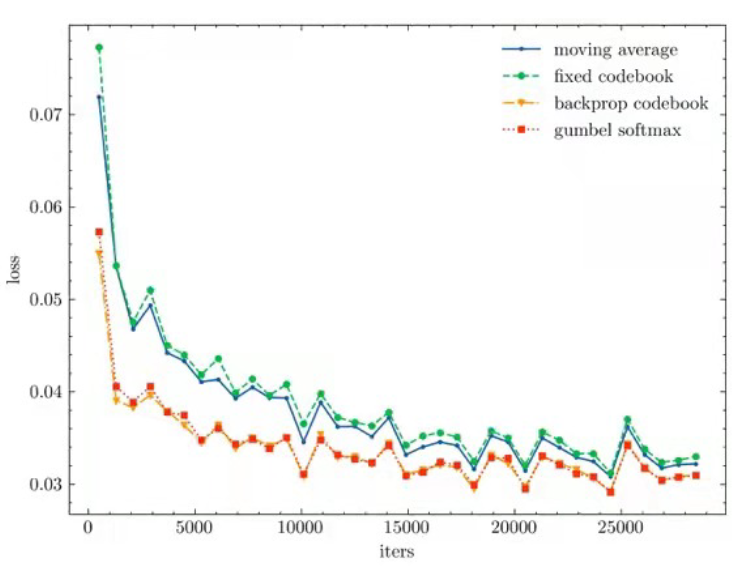


图2：训练图像标记器时的L2损失曲线。上述所有的方法最终都收敛到一个类似的损失水平。

我们发现，所有的方法基本上都是势均力敌的，这意味着如果初始化得当，编码本中的嵌入学习并不是很重要。在预训练中，我们使用移动平均法的标记器。数据的介绍和关于标记化的更多细节见附录A。

2.3 自动回归变换器  
CogView的骨干是一个单向转化器（GPT）。变换器有48层，隐藏大小为2560，40个注意头，总共有40亿个参数。如图3所示，

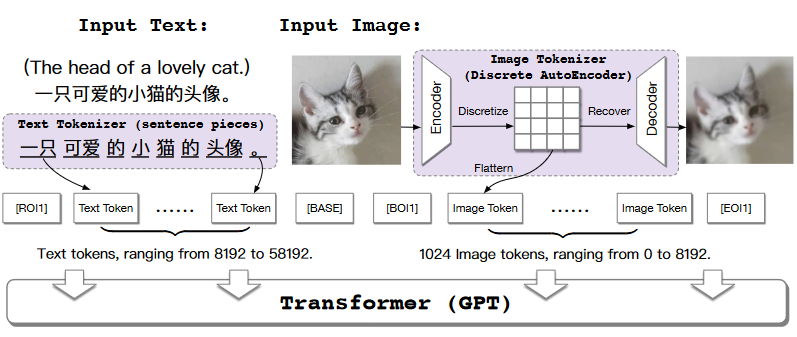


图3：CogView的框架。[ROI1]、[BASE1]等，是分隔符号。

四个分隔符，[ROI1]（图像的参考文本），[BASE]，[BOI1]（图像的开始），[EOI1]（图像的结束）被添加到每个序列中，以指示文本和图像的边界。所有的序列都被剪掉或填充到1088的长度。预训练的前文任务是从左到右的标记预测，又称语言建模。图像和文本令牌都被同等对待。DALL-E[39]建议降低文本标记的损失权重；相反，在小规模的实验中，我们令人惊讶地发现文本建模是文本到图像预训练成功的关键。如果文本标记的损失权重被设置为零，模型将无法找到文本和图像之间的联系，并产生与输入文本完全无关的图像。  
我们假设，文本建模在隐蔽层中抽象出知识，这些知识可以在后面的图像建模过程中被有效利用。我们在512个V100 GPU（32GB）上以6144个序列（每批670万个标记）的批量大小训练模型，进行144,000步训练。参数由Adam更新，最大lr = 3 × 10 -4 ,β 1 = 0.9, β 2 = 0.95, 权重衰减 = 4 × 10 -2 。学习率在最初的2%步骤中升温，并随着余弦退火法[34]衰减。在超参数处于适当范围的情况下，我们发现训练损失主要取决于训练的标记总数（每批标记×步骤），这意味着如果训练相同数量的标记，批次大小（和学习速率）增加一倍会导致非常相似的损失。因此，我们使用一个相对较大的批处理量来提高并行性，减少通信时间的百分比。我们还设计了一个三区稀疏注意，以加快训练速度，节省内存，而不损害性能，这在附录B中介绍。

2.4 训练的稳定性

目前，预训练大型模型（>2B参数）通常依靠16位精度来节省GPU内存并加快计算速度。许多框架，例如DeepSpeed ZeRO[40]，甚至只支持FP16参数。然而，文本到图像的预训练在16位精度下是非常不稳定的。训练一个4B的普通预LN变换器，在1000次迭代内很快就会出现NaN损失。要稳定训练是CogView最具挑战性的部分，它与DALL-E有很好的一致性。我们总结了DALL-E的解决方案，以容忍训练的数字问题。由于不同层的值和梯度在规模上有很大的不同，他们提出了一个新的混合精度框架，每块损失缩放，并以32位精度存储所有的增益、偏置、嵌入和非嵌入，并有32位梯度。这个解决方案很复杂，需要消耗额外的时间和内存，并且不被当前大多数训练框架所支持。CogView代替了对数值的规范化。我们发现有两种不稳定性：溢出（以NaN损失为特征）和下溢（以分歧损失为特征）。我们提出了以下技术来解决它们。

**精度瓶颈放松法（PB-Relax）**。在分析了训练的动态后，我们发现溢出总是发生在两个瓶颈操作，即最后的LayerNorm或注意力。

- 在深层中，输出的值可能会爆发到10 4∼10 5，使得LayerNorm的变化发生溢出。幸运的是，由于LayerNorm(x)=LayerNorm(x/ max(x))，我们可以通过先除以最大值4来放松这个瓶颈。

- 注意力分数Q T K/ √ d可能明显大于输入元素，并导致溢出。将计算顺序改为Q T (K/ √ d)可以缓解这个问题。为了消除溢出，我们注意到softmax(Q T K/ √ d) = softmax(Q T K/ √ d-constant)，也就是说，我们可以将注意力的计算改为softmax( QT K √d ) = softmax ( ( QT α√d K - max( QT α√d K) ) × α ) , (4) 其中α是一个大数，例如α=32.5 这样一来，注意力分数的最大值（绝对值）也被α除以，以防止其溢出。关于CogView的注意力的详细分析见附录C。

三明治层规范（Sandwich-LN）。Transformer中的LayerNorms[1]对于稳定的训练至关重要。Pre-LN[50]被证明比原来的Post-LN收敛得更快、更稳定，在最近的工作中成为Transformer层的默认结构。然而，对于文本到图像的预训练来说，它是不够的。LayerNorm (x- ̄ x)√d √ ∑ i (xi - ̄ x)2 γ + β的输出基本上与x的隐藏大小的平方根成正比，在CogView中是√d = √2560≈50。如果某些维度的输入值明显大于其他维度--对于变形金刚来说就是这样--这些维度的输出值也会很大（101 ∼ 102）。在剩余分支中，这些大值被放大并被加回主分支，这在下一层中加剧了这种现象，最终导致深层中的值爆炸。这种价值爆炸背后的原因激发了我们对逐层加重现象的限制。我们提出了Sandwich LayerNorm，它还在每个剩余分支的末端添加了一个LayerNorm。SandwichLN保证了每层输入值的尺度在一个合理的范围内，对500M模型的训练实验表明，它对收敛的影响可以忽略不计。图4(a)说明了变形金刚中不同的LayerNorm结构。玩具实验。图4(b)显示了PB-relax和Sandwich-LN在玩具实验中的效果，因为训练许多大型模型进行验证是不现实的。我们发现，深层变压器（64层，1024个隐藏大小），大的学习率（0.1或0.01），小的批次大小（4）可以在合理的超参数下模拟训练中的价值爆炸。PB-relax + Sandwich-LN甚至可以稳定玩具实验。缩小嵌入梯度。虽然我们在使用Sandwich-LN后没有观察到任何下溢的迹象，但我们发现令牌嵌入的梯度比其他参数的梯度大得多，因此，只要将其规模缩小α=0.1，就可以增加动态损失规模，进一步防止下溢，这可以通过Pytorch中的emb=emb\*alpha+emb.detach()\*(1-alpha)实现。这似乎减慢了标记嵌入的更新速度，但实际上在我们的实验中并没有损害性能，这也与最近的工作MoCo v3[9]相一致。讨论。PB-relax和Sandwich-LN成功地稳定了CogView和一个8.3B参数的CogView-large的训练。它们对于所有的变形器预训练也是通用的，并将在未来实现对非常深的变形器的训练。作为一个证据，我们使用PB-relax成功地消除了训练10B参数的GLM的溢出[14]。然而，一般来说，语言预训练中的精度问题并不像文本到图像预训练中那么显著。我们假设其根源在于数据的异质性，因为我们观察到文本和图像标记在某些隐藏状态下是有尺度区分的。另一个可能的原因是难以发现的下溢，由DALL-E猜测的。彻底的调查将留给未来的工作。

3 微调

CogView在微调方面比DALL-E走得更远。特别是，我们可以通过微调CogView的超分辨率和自我排序来改善文本到图像的生成。所有的微调任务都可以在一天之内在一台DGX-2上完成。

3.1 超级分辨率

由于图像标记器在训练前将256×256像素的图像压缩成32×32的标记序列，由于有损压缩，生成的图像比真实图像更模糊。然而，由于注意力操作的O(n2)的复杂性，扩大序列长度将消耗更多的计算和内存。以前关于超分辨率或图像修复的工作[13]，通常处理已经是高分辨率的图像，将模糊的局部纹理映射到清晰的纹理。它们不能适用于我们的情况，我们需要在生成的低分辨率图像中添加有意义的细节。图5（b）是我们微调方法的一个例子，说明了我们所期望的超分辨率行为。

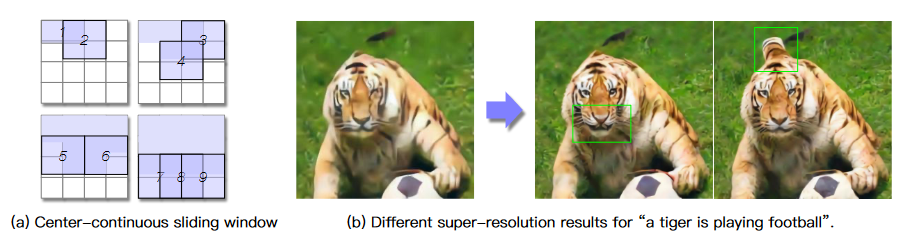


图5：（a）一个64×64的令牌图像按照数字顺序逐个生成补丁。重叠的位置将不会被覆盖。关键的想法是使第2和第4个区域的令牌--通常是人脸区域或其他重要部分--在关注整个区域时生成。(b) 经过微调的超级分辨率模型并不勉强改造纹理，而是生成新的局部结构，例如例子中的张开的嘴或尾巴。

我们对超分辨率进行微调的动机是相信CogView是在一般领域中最复杂的分布上训练出来的，而且不同分辨率的物体已经被覆盖了。具体来说，我们首先将CogView微调为一个有条件的超分辨率模型，从16×16的图像标记到32×32的标记。然后，我们通过图5（a）中的中心连续滑动窗口策略将32×32标记的图像逐片放大到64×64标记（512×512像素）。这种顺序在保持中心区域的完整性方面比栅格扫描顺序表现得更好。为了准备数据，我们将大约200万张图像裁剪成256×256的区域，并将其降样为128×128。在标记化之后，我们得到不同分辨率的32×32和16×16的序列对。微调序列的模式是"[ROI1]文本标记[BASE][BOI1] 16×16图像标记[EOI1] [ROI2][BASE] [BOI2] 32×32图像标记[EOI2]"，比最大位置嵌入索引1087长。作为一个解决方案，我们从0开始重新计算[ROI2]的位置索引。

3.2 图像标题和自我排序

对CogView的图像标题进行微调是很简单的：交换输入序列中的文本和图像标记的顺序。由于该模型已经学会了文本和图像之间的相应关系，反转生成并不困难。我们没有评估性能，因为（1）没有权威的中文图像字幕基准（2）图像字幕不是这项工作的重点。对这样一个模型进行微调的主要目的是为了进行自我排序。我们提出标题损失（CapLoss）来评估图像和文本之间的对应关系。更具体地说，CapLoss(x, t) = 1 |t| ∑|t| i=0 - log p(ti|x, t0:i-1)，其中t是文本标记的序列，x是图像。CapLoss(x, t)是文本标记的交叉熵损失，这种方法可以看作是对文本到图像生成的逆向提示法[56]的一种改编。最后，选择具有最低CapLosses的图像。与额外训练另一个约束性自我监督模型，例如CLIP[38]，用于重新排序相比，我们的方法消耗的计算资源更少，因为我们只需要微调。图9中的结果显示，用我们的方法选择的图像在FID中比用CLIP选择的图像表现得更好。图6显示了一个重新排序的例子。



图6："一个穿红衬衫的人在玩电子游戏 "的60张生成的图片（从COCO中随机选择），按CapLoss的顺序显示。大多数糟糕的案例都排在最后。这种多样性也缓解了人们对CogView可能对训练集中的类似图像过度拟合的担忧。

3.3 风格学习

虽CogView经过预训练，能够尽可能地覆盖各种图像，但不能很好地满足生成特定风格或主题的图像的愿望。我们对四种风格的模型进行了微调。中国传统绘画、油画、素描和卡通。这些风格的图片都是从包括谷歌、百度和必应等搜索引擎页面中自动提取的，关键词为"{style}风格的图片"，其中{style}是风格的名称。我们分别对不同风格的图片进行微调，每张图片有1,000张。在微调过程中，图像的相应文字也是"{style}风格的图像"。生成时，文本是 "A {object} of {style} style"，其中{object}是要生成的对象。通过这种方式，CogView可以将从预训练中学到的对象的形状知识转移到微调的样式中。图7显示了各种样式的例子。

3.4 工业时尚设计

当生成的目标是单一领域时，纹理的复杂性就会大大降低。在这些情况下，我们可以（1）训练一个VQGAN[15]，而不是VQVAE的潜变量，以获得更真实的纹理；（2）减少参数的数量，增加序列的长度，以获得更高分辨率。我们的三区稀疏注意（附录B）在这种情况下可以加快高分辨率图像的生成。我们在大约1000万个时尚标题对上训练一个3B参数模型，使用50×50的VQGAN图像标记，并将它们解码成800×800像素。图8显示了CogView用于时装设计的样本，它已经成功部署到阿里巴巴Rhino时装生产中。

5 结论和讨论  
**局限性** CogView的缺点是生成速度慢，这对自动回归模型来说是很常见的，因为每张图像都是逐个令牌生成。VQVAE带来的模糊性也是一个重要的限制。这些问题将在未来的工作中得到解决。  
**伦理方面的关注** 与Deepfake类似，CogView也很容易被恶意使用[49]，因为它有可控的、强大的生成图像的能力。缓解这一问题的可能方法在一项调查中讨论过[5]。此外，在关于人类9的生成模型中通常存在公平性问题。在附录D中，我们分析了CogView的公平性情况，并介绍了一个简单的 "单词替换 "方法来解决这个问题。  
我们系统地研究了结合VQVAE和Transformers进行文本到图像生成的框架。CogView为可扩展的跨模式生成预训练展示了有希望的结果，同时也揭示并解决了可能源于数据异质性的精度问题。我们还介绍了为不同的下游任务对CogView进行微调的方法。我们希望CogView能够推动可控图像生成和跨模态知识理解的研究和应用，但需要防止它被用于创建错误信息的图像。