女娲：神经视觉世界创建的视觉合成预训练

陈飞吴江梁磊纪梵杨月鉴房大新江南段1\* 2\* 1            1            2            1            1†

1微软亚洲研究院2北京大学

{周、雷吉、范阳、蒋、南端}@microsoft。com{j。liang@stu,fangyj@ss}北大校长。埃杜。cn

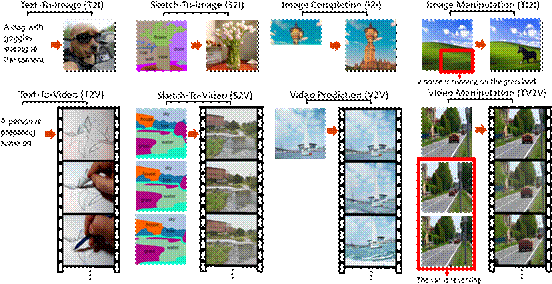


图1。NUWA模型支持的8个典型视觉生成和操作任务的示例。¨

摘要

本文提出了一种称为NUWA的统一多模态预训练模型，该模型可以为各种视觉合成任务生成新的或操作现有的视觉数据（即图像和视频）。为了在不同场景下同时覆盖语言、图像和视频，设计了一个3D transformer编解码框架，该框架不仅可以将视频作为3D数据处理，还可以分别将文本和图像作为一维和二维数据处理。还提出了一种3D邻近注意力（3DDNA）机制来考虑视觉数据的性质，并降低计算复杂度。我们在8项下游任务中评估女娲。与几个强大的基线相比，NUWA？在文本到图像生成、文本到视频生成、视频预测等方面取得了最先进的成果。此外，它在文本引导的图像和视频操作任务方面也表现出令人惊讶的良好零拍能力。回购计划https://github.com/微软/女娲。



\*两位作者对这项研究的贡献相同。

†通讯作者。

# 1.导言

如今，随着图像和视频成为新的信息载体并在许多实际应用中得到应用，网络变得比以往任何时候都更加可视化。在这样的背景下，视觉合成正成为一个越来越热门的研究课题，其目的是为各种视觉场景构建能够生成新的或操作现有视觉数据（即图像和视频）的模型。

自回归模型[33,39,41,45]在视觉合成任务中发挥着重要作用，因为与GANs相比，自回归模型具有明确的密度建模和稳定的训练优势[4,30,37,47]。早期的视觉自回归模型，如PixelCNN[39]、PixelRNN[41]、图像转换器[28]、iGPT[5]和视频转换器[44]，以“逐像素”的方式执行视觉合成。然而，由于其对高维视觉数据的高计算成本，此类方法只能应用于低分辨率图像或视频，并且难以扩展。

最近，随着VQ-VAE[40]作为一种离散视觉标记化方法的出现，高效和大规模的预处理-

1

培训可应用于图像（如DALL-e[33]和CogView[9]）和视频（如GODIVA[45]）的视觉合成任务。尽管取得了巨大的成功，但这些解决方案仍有局限性——它们将图像和视频分开处理，并专注于生成其中任何一种。这限制了模型同时受益于图像和视频数据。

在本文中，我们提出了NUWA，一个统一的多模态预训练模型，旨在支持图像和视频的视觉合成任务，并在8个下游视觉合成上进行了实验，如图1所示。这项工作的主要贡献有三个方面：

•我们提出NUWA，一个通用的3D transformer编码器-解码器框架，它同时涵盖语言、图像和视频，用于不同的视觉合成任务。它由一个以文本或视觉草图为输入的自适应编码器和一个由8个视觉合成任务共享的解码器组成。

我们在框架中提出了一个3D附近注意力（3DNA）机制来考虑空间和时间轴的局部特性。3DNA不仅降低了计算复杂度，而且提高了生成结果的视觉质量。

•与几个强大的基线相比，NUWA在文本到图像生成、文本到视频生成、视频预测等方面取得了最先进的成果。此外，NUWA不仅在文本引导的图像处理方面，而且在文本引导的视频处理方面，表现出令人惊讶的良好零拍能力。

# 2.相关工作

## 2.1. 可视化自回归模型

本文提出的方法遵循了基于自回归模型的视觉合成研究路线。早期的视觉自回归模型[5,28,39,41,44]以“逐像素”的方式进行视觉合成。然而，由于高维数据建模时计算成本高，此类方法只能应用于低分辨率图像或视频，并且难以扩展。

最近，基于VQ VAE的[40]视觉自回归模型被提出用于视觉合成任务。通过将图像转换为离散的视觉标记，此类方法可以对文本到图像生成（例如，DALL-e[33]和CogView[9]）、文本到视频生成（例如，GODIVA[45]）和视频预测（例如，LVT[31]和VideoGPT[48]）进行高效和大规模的预训练，生成的图像或视频具有更高的分辨率。然而，这些模型都不是通过图像和视频一起训练的。但直观的是，这些任务可以从这两种类型的可视数据中获益。

与这些作品相比，NUWA是一个统一的自动回归视觉合成模型，由覆盖图像和视频的视觉数据预先训练，可以支持各种下游任务。我们还验证了Sec中不同预训练任务的有效性。4.3. 此外，在女娲使用VQ-GAN[11]代替VQ-VAE进行语音标记化，根据我们的实验，这可以提高生成质量。

## 2.2. 视觉稀疏自我注意

如何处理自我关注带来的二次复杂性问题是另一个挑战，特别是对于高分辨率图像合成或视频合成等任务。

与NLP类似，稀疏注意机制也被用来缓解视觉合成中的这个问题。[31,44]将视觉数据分割成不同的部分（或块），然后对合成任务执行分块稀疏注意。然而，这些方法分别处理不同的块，并且没有对它们的关系建模。[15,33,45]提出在视觉合成任务中使用轴向稀疏注意，该任务沿着视觉数据表示的轴进行稀疏注意。这种机制使得训练非常有效，并且对大规模预训练模型（如DALL-E[33]、CogView[9]和GODIVA[45]）非常友好。然而，由于自我注意中使用的语境有限，生成的视觉内容的质量可能会受到损害。[6,28,32]建议在视觉合成任务中使用局部稀疏注意，这使得模型能够看到更多的上下文。但这些作品只是为了图像。

与这些作品相比，女娲提出了一种3D近在咫尺的注意力，它扩展了局部的稀疏注意力，将两幅图像都覆盖到视频中。我们还验证了局部稀疏注意比轴向稀疏注意更适用于Sec中的视觉生成。4.3.

# 3.方法

## 3.1. 三维数据表示

为了涵盖所有文本、图像和视频或它们的草图，我们将它们视为标记并定义统一的3D符号∈ R、 其中和表示空间轴上的令牌数量（分别为高度和宽度），表示时间轴上的令牌数量，是每个令牌的维度。在下面，我们将介绍如何获得不同模式的统一表示。*十、H*×w×s×d*HWsD*

文本自然是离散的，在Transformer[42]之后，我们使用小写字节对编码（BPE）来标记它们并将它们嵌入到R中。我们使用占位符1，因为文本没有空间维度。1×1×s×d

图像自然是连续的像素。输入原始图像∈ R根据高度、宽度和通道，VQ-VAE[40]训练一个可学习的码本，在原始连续像素和离散标记之间建立桥梁，如de-*我H*×W×C*HWC*

|  |
| --- |
| 图2。概述女娲的结构。它包含一个支持不同条件的自适应编码器和一个预先训练的解码器，可以从图像和视频数据中获益。对于图像完成、视频预测、图像处理和视频处理任务，输入 |

部分图像或视频直接输入解码器。

如式（1）所示∼(2):

*梓*=argmin | E（I）I− 北京| | 2，(1)

*J*

*我*ˆ=G（B[z]），(2)

其中是编码为×w网格特征的编码器（I）∈ R∈ Ris是一个具有可视标记的可学习代码本，其中（I）的每个网格都会被搜索以找到最近的标记。搜索结果∈ {0,1，…，N− 1} h×w由解码器嵌入并重构回。VQ-VAE的训练损失可写成公式（3）：*E我HEH*×w×d*BBN*×d*BNEZB我*ˆ *G*

*吕*=| |我−I | | 22+| | sg[E（I）]−B[z]| | 22+| | E（I）−sg[B[z]| | 22，ˆ

(3)

其中严格限制了两者之间的精确像素匹配*我*这限制了模型的泛化能力。最近，VQ-GAN[11]通过添加感知损失和GAN损失来增强VQ-VAE训练，以缓解和之间的精确约束，并专注于高级语义匹配，如等式（4）所示∼(5):*我*ˆ*我我*ˆ

*,* (4)

|  |  |
| --- | --- |
| *LG*=logD（I）+log1− D（I））。ˆ | (5) |

VQ-GAN培训后，[z]∈ Ris最终被用作图像的表示。我们使用占位符1，因为图像没有时间维度。*BH*×宽×1×d

视频可以被视为图像的时间扩展，最近的工作，如VideoGPT[48]和VideoGen[51]将VQ-VAE编码器中的卷积从2D扩展到3D，并训练视频特定的表示。但是，这无法为图像和视频共享一个通用的代码本。在本文中，我们证明了简单地使用2D VQ-GAN对视频的每一帧进行编码也可以生成时间一致性视频，同时受益于图像和视频数据。结果表示表示为R，其中表示帧数。*H*×w×s×d*s*

对于图像素描，我们认为它们是具有特殊通道的图像。图像分割矩阵rw的每个值表示一个像素的类别，可以以一种热的方式查看rw，这里是分割类别的数量。通过为图像草图训练额外的VQGAN，我们最终得到嵌入的图像表示R。同样，对于视频草图，表示为。*H*×W*H*×W×C*CH*×宽×1×d*相对湿度*×w×s×d

## 3.2. 三维近距离自我注意

在本节中，我们基于先前的3D数据表示定义了一个统一的3D邻近自我注意（3DNA）模块，支持自我注意和交叉注意。我们首先在公式（6）中给出3DNA的定义，并在公式（7）中介绍详细的实现∼(11):

*Y*=3DNA（X，C；W），(6)

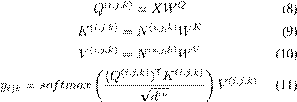
两者都在哪里∈ R×w×s×d∈ 第节中介绍的R0×w0×s0×dare 3D表示法。3.1. 如果=X，3DNA表示对目标的自我注意，如果6=X，3DNA表示对目标的交叉注意。表示可学习的权重。*十、H在里面CH在里面C十、C十、CW*

我们开始从下面的坐标（i，j，k）引入3DNA。通过线性投影，对应的坐标（i0，j0，k0）*十、*在下面然后，附近的当地居民（i0，j0，k0）具有宽度、高度和时间范围∈ 式（7）中定义的Ris，*嗯,嗯,嗯*+

*,*

其中（i，j，k）∈ R×e×e×dis是条件的次张量，由（i，j，k）需要注意的相应邻近信息组成。有三个可学习的权重*NEHWs在里面C*

*WQ，WK，WV*∈ R、 位置（i，j，k）的输出张量如式（8）所示∼(11):*D在里面*×d*出来*



其中（i，j，k）位置查询并收集中相应的附近信息。这也处理=X，然后（i，j，k）*CC*只需查询其自身的附近位置。3NDA不仅降低了从到的全部注意力的复杂性，而且显示了优异的性能，我们将在第。4.3.

## 3.3. 三维编解码器

在本节中，我们将介绍基于3DNA构建的3D编码解码器。生成目标∈ R*YH*×w×s×d*出来*

在∈ R、 和的位置编码由三个不同的可学习词汇更新，分别考虑高度、宽度和时间轴，如式（12）所示∼(13):*CH*0×w0×s0×d*在里面YC*

*Yijk*：=Y*ijk*+P*ih*+P*jw*+P*ks*(12)

                                                        0                  0                0

*Cijk*：=C*ijk*+P*ih*+P*jw*+P*ks*(13)

然后，将该条件输入到具有3DNA层堆栈的编码器中，以对自我注意交互进行建模，其中第th层在等式（14）中表示：*CLL*

*C*（l）=3DNA（C）（l）−1)*C*（l）−1)), (14)

类似地，解码器也是3DNA层的堆栈。解码器计算生成结果的自我注意以及生成结果和条件之间的交叉注意。第th层在等式（15）中表示。*LL*

                       （l） （l）−1） （l）−1)

*Yijk*=3DNA（Y）*<i，<j，<kY<i，<j，<k*

(15)

                                                   （l）−1） （L）

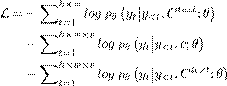
                                  +3DNA（Y），*<i，<j，<kC*

其中表示目前生成的令牌。*<i，<j，<k*

初始标记是一个特殊的标记*<bos>*在培训阶段学习的代币。

## 3.4. 培训目标

我们在三个任务上训练我们的模型：文本到图像（T2I）、视频预测（V2V）和文本到视频（T2V）。这三项任务的培训目标是交叉熵，分别在公式（16）中表示为三个部分：

(16)

对于T2I和T2V任务，表示文本条件。对于V2V任务，由于没有文本输入，我们得到了特殊单词“None”的恒定三维表示。表示模型参数。*CtextCθ*

# 4.实验

根据美国证券交易委员会。3.4我们首先在三个数据集上对NUWA进行预训练：用于文本到图像（T2I）生成的概念性标题[22]，其中包括290万个文本-图像对；用于视频预测（V2V）的时间矩[26]，其中包括727K个视频；以及用于文本到视频（T2V）生成的VATEX数据集[43]，其中包括241K个文本-视频对。在下文中，我们首先在第2节中介绍实现细节。4.1然后将NUWA与第节中的最先进模型进行比较。4.2，最后在Sec中进行烧蚀研究。4.3研究不同部分的影响。

## 4.1. 实施细节

以秒计。3.1，我们将文本、图像和视频的3D表示大小设置如下。对于文本，三维表示的大小为1×1×77×1280。对于图像，三维表示的大小为21×21×1×1280。对于视频，3D表示的大小为21×21×10×1280，其中我们从一个每秒2.5帧的视频中采样10帧。虽然默认的视觉分辨率为336×336，但我们预先训练了不同的分辨率，以便与现有模型进行公平比较。对于用于图像和视频的VQ-GAN模型，等式（1）中网格特征（I）的大小为441×256，码本的大小为12288。*EB*

不同的稀疏范围用于Sec中的不同模式。3.2. 对于文本，我们设置（ew，eh，es）=（1,1，∞), 哪里∞ 表示在注意时始终使用全文。对于图像和图像草图，（ew，eh，es）=（3,3,1）。对于视频和视频草图，（ew，eh，es）=（3,3,3）。

我们在64个A100 GPU上进行为期两周的预训练，将等式（14）中的图层设置为24，这是一个Adam[17]优化器，学习率为1e-3，批量大小为128，预热总步骤为50M的5%。最终预训练模型的参数总数为870M。*L*

一只非常可爱的猫中国航空公司普通的一张桌子有一个客厅和两个人一只非常可爱的长颈鹿一个厨房有一群动物躺在一辆大自行车旁边。地面上有一个火车模型，上面有一台电视机，上面有一个正在做有趣的冰箱、炉子，站在机场里拿着行李和其他汽车，站在一个吉他木长凳上。面对下沉雪附近有汽车。等等。坐在他旁边。



一列绿色的火车来了，一群滑雪者在一个小厨房，一个起居室，一个孩子在气球旁吃生日蛋糕。沿着轨道往下走。准备用低天花板滑雪。电视和桌子。

                                   下山。XMC-GAN（256×256）NÜWA（我们的）（256×256）

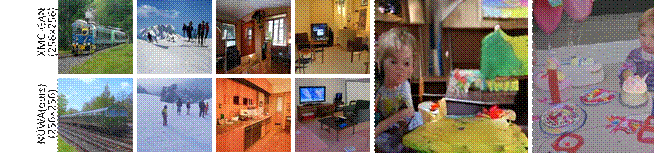
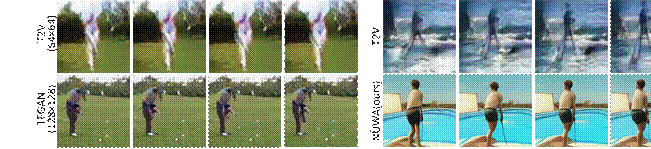


图3。与MSCOCO数据集上文本到图像（T2I）任务的最新模型进行定性比较。

          输入文本：在草地上打高尔夫球输入文本：在游泳池打高尔夫球



输入文本：在海上跑步



图4。动力学数据集上文本到视频（T2V）任务与最新模型的定量比较。

4.2. 与最先进的ing GODIVA[45]相比，我们使用了CLIPSIM度量，其-

文本到图像（T2I）微调：我们将NUWA？porates与CLIP[29]模型进行比较，以计算输入文本和生成图像之间的语义相似性。对于MSCOCO[22]数据集上的展会，请参见表1。1与之相比，所有模型的分辨率均为256×256。定性地，如图3所示。根据DALL-E[33]，我们使用

“初始值”和“最佳分数”分别为[60分和[35分]（我们为每个FID图像生成一个分数）[初始值]和[35分]。在选项卡中。1、女娲在质量和品种方面明显优于–分别进行评估，并遵循-*KK*

表1。与MSCOCO（256×256）数据集上文本图像转换（T2I）任务的最新模型进行定性比较。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | FID-0↓ | FID-1 FID-2 FID-4 FID-8为↑ | | | 阴蒂 |
| AttnGAN[47] | 35.2 | 44.0 | 72.0 | 108.0 100.0 23.3 | 0.2772 |
| DM-GAN[52] | 26.0 | 39.0 | 73.0 | 119.0 112.3 32.2 | 0.2838 |
| DF-GAN[36] | 26.0 | 33.8 | 55.9 | 91.0    97.0       18.7 | 0.2928 |
| DALL-E[33] | 27.5 | 28.0 | 45.5 | 83.5    85.0       17.9 | - |
| CogView[9] | 27.1 | 19.4 | 13.9 | 19.4    23.6       18.2 | 0.3325 |
| XMC-GAN[50] | 9.3 | - | - | -          -            30.5 | - |
| 女娲 | 12.9 | 13.8 | 15.7 | 19.3    24         27.2 | 0.3429 |

↑

表2。动力学数据集上文本到视频（T2V）任务与最新模型的定量比较。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 行政协调会↑ | 氢火焰离子化↓ | 氢火焰离子化检测器↓ | 阴蒂↑ |
| T2V（64×64）[21] | 42.6 | 82.13 | 14.65 | 0.2853 |
| SC（128×128）[2] | 74.7 | 33.51 | 7.34 | 0.2915 |
| TFGAN（128×128）[2] | 76.2 | 31.76 | 7.19 | 0.2961 |
| 女娲（128×128）（ | 77.9 | 28.46 | 7.05 | 0.3012 |

表3。在BAIR（64×64）数据集上与最先进的视频预测（V2V）任务模型进行定量比较。

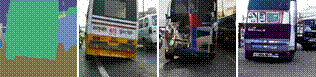
|  |  |  |
| --- | --- | --- |
| 模型 | 康德。 | FVD↓ |
| 摩卡咖啡[37] | 4 | 503 |
| SVG-FP[8] | 2 | 315 |
| 中国国家数据局[12] | 2 | 297 |
| SV2P[1] | 2 | 263 |
| SRVP[13] | 2 | 181 |
| 视频流[18] | 3 | 131 |
| LVT[31] | 1 | 126±3 |
| SAVP[20] | 2 | 116 |
| DVD-GAN-FP[7] | 1 | 110 |
| 视频转换器[44] | 1 | 106±3 |
| TriVD-GAN-FP[23] | 1 | 103 |
| CCVS[25] | 1 | 99±2 |
| 视频转换器（L）[44] | 1 | 94±2 |
| 女娲 | 1 | 86.9 |

CogView[9]，FID-0为12.9，CLIPSIM为0.3429。尽管XMC-GAN[50]报告的FID分数为9.3，但我们发现，与XMC-GAN论文中完全相同的样本相比，女娲生成的图像更真实（见图3）。特别是在最后一个示例中，男孩的脸是清晰的，气球是正确生成的。

文本到视频（T2V）微调：我们在Tab中定量比较动力学[16]数据集上的NUWA。2和图4中的定性。根据TFGAN[2]，我们评估FID img和FID vid度量的视觉质量以及生成视频标签准确性的语义一致性。如选项卡中所示。2、女娲在上述所有指标上都取得了最佳绩效。在图4中，我们还展示了生成不可见文本的强大零拍能力，例如“在游泳池打高尔夫”或“在海上跑步”。

视频预测（V2V）微调：我们在选项卡中定量比较了BAIR机器人推送[¨10]数据集上的NUWA。3.康德。表示给定给的帧数

输入地面真相黑桃（256×256）驯服（256×256）

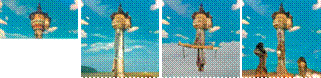
NÜWA（我们的）

NÜWA（我们的）（256×256）



图5。与MSCOCO stuff数据集上草图到图像（S2I）任务的最新模型进行定量比较。

输入驯化（256×256）

NÜWA（我们的）

NÜWA（我们的）（256×256）

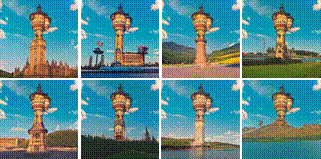


图6。以零拍方式与最先进的图像完成（I2I）任务模型进行定性比较。

预测未来的框架。为了公平比较，所有模型都使用64×64分辨率。尽管只给出一帧作为条件（Cond.），但，女娲仍然显著地将最先进的FVD[38]得分从94±2提高到86.9。

草图到图像（S2I）微调：我们在图5中定性地比较了NUWA与MSCOCO材料[¨22]。与驯服变形金刚[11]和斯派德[27]相比，女娲能生产出种类繁多的逼真巴士。甚至公共汽车车窗的反光也清晰可见。

图像完成（I2I）零拍评估：我们在图6中以零拍的方式定性地比较女娲。与驯服变形金刚[11]相比，女娲在塔楼的上半部分展示了对塔楼下半部分的丰富想象，包括建筑物、湖泊、鲜花、草地、树木、山脉等。

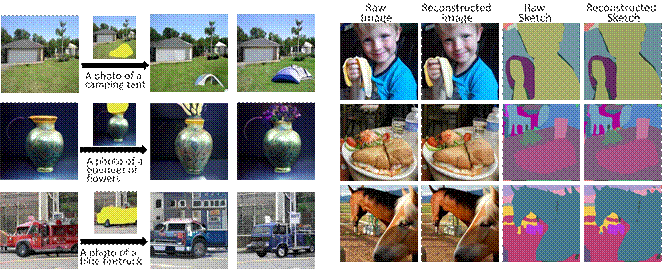
单词NÜWA（我们的）的原始图像处理绘画

图7。以零拍方式与最先进的文本引导图像处理（TI2I）模型进行定量比较。

文本引导图像处理（TI2I）零镜头评估：我们以零镜头方式比较NUWA–定性如图7所示。与单词绘制[3]相比，NUWA显示出强大的操作能力，生成高质量的文本一致性结果，同时不改变图像的其他部分。例如，在第三排，女娲产生的蓝色消防车更为逼真，而后面的建筑则没有变化。这得益于通过对各种视觉任务进行多任务预训练而获得的真实视觉模式。另一个优势是女娲的推理速度，几乎50秒就可以生成一个图像，而文字绘画在推理过程中需要额外的训练，大约需要300秒才能收敛。

草图到视频（S2V）微调和文本引导视频操作（TV2V）零镜头评估：据我们所知，开放域S2V和TV2V是本文首次提出的任务。由于没有比较，我们将其安排在第4.3节的消融研究中。

附录中提供了更详细的比较、样本，包括人类评估。

## 4.3. 烧蚀研究

选项卡的上面部分。图4显示了不同VQ-VAE（VQ-GAN）设置的有效性。OpenNet和ImageWe实验[34]。表示原始分辨率，表示离散标记的数量。压缩率表示为√ √ Fx，其中是*RD十、*

*R*除以。比较选项卡中的前两行。与VQ-VAE相比，VQ-GAN显示出明显更好的Frechet初始偏差（FID）[14]和结构相似矩阵（SSIM）得分。比较第2-3行，我们发现离散标记的数量是导致更高视觉质量的关键因素，而不是压缩率。虽然第2排*D*

图8。VQ-GAN和VQ-GAN Seg的重建样品。表4。不同VQ-VAE（VQ-GAN）设置的有效性。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 数据集 | *R*→D | 速率SSIM | | 氢火焰离子化检测器 |
| VQ-VAE | 图像网 | 2562→ 162 | F16 | 0.7026 | 13.3 |
| VQ-GAN | 图像网 | 2562→ 162 | F16 | 0.7105 | 6.04 |
| VQ-GAN | 图像网 | 2562→ 322 | F8 | 0.8285 | 2.03 |
| VQ-GAN | 图像网 | 3362→ 212 | F16 | 0.7213 | 4.79 |
| VQ-GAN | OpenImages | 3362→ 212 | F16 | 0.7527 | 4.31 |
| 模型 | 数据集 | *R*→D | 费率PA | | FWIoU |
| VQ-GAN-Seg | MSCOCO | 3362→ 212 | F16 96.82 | | 93.91 |
| VQ-GAN-Seg | VSPW | 3362→ 212 | F16 95.36 | | 91.82 |

第4行具有相同的压缩率F16，它们的FID分数分别为6.04和4.79。因此，重要的不仅仅是我们压缩了多少原始图像，还有多少离散标记用于表示图像。这与认知逻辑是一致的，仅仅用一个标记来表示人脸太模糊了。实际上，我们发现162个离散标记通常会导致较差的性能，尤其是人脸，322个标记显示出最好的性能。然而，更多的离散代币意味着更多的计算，特别是对于视频。我们最终使用了一个折衷版本进行预培训：212个代币。通过在开放图像数据集上进行训练，我们进一步将212版本的FID分数从4.79提高到4.31。

选项卡的下面部分。图4显示了用于草图的VQGAN的性能。MSCOCO[22]上的VQ GAN Seg接受了草图到图像（S2I）任务和VQ GAN Seg的培训

在VSPW[24]上进行了草图到视频（S2V）任务的培训。所有这些主干在像素精度（PA）和联合上的频率加权交集（FWIoU）方面表现出良好的性能，这表明我们的模型中使用的三维草图表示具有良好的质量。图8还显示了336×336图像和草图的一些重建样本。

表5。MSRVTT数据集上文本到视频（T2V）生成任务多任务预训练的有效性。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 预先培训的任务 | 氢火焰离子化检测器↓ | 阴蒂↑ |
| 女娲电视台 | T2V | 52.98 | 0.2314 |
| 女娲电视台 | T2V+T2I | 53.92 | 0.2379 |
| NUWA-TV-VV¨ | T2V+V2V | 51.81 | 0.2335 |
| 女娲 | T2V+T2I+V2V | 47.68 | 0.2439 |

标签。5显示了文本到视频（T2V）生成任务的多任务预训练的有效性。我们研究了一个具有挑战性的数据集，MSR-VTT[46]，其中包含自然描述和真实视频。与仅针对单个T2V任务（第1行）的培训相比，针对T2V和T2I（第2行）的培训将CLIPSIM从0.2314提高到0.2379。这是因为T2I有助于在文本和图像之间建立连接，从而有助于T2V任务的语义一致性。相比之下，对T2V和V2V（第3行）的训练将FVD得分从52.98提高到51.81。这是因为V2V有助于学习常见的无条件视频模式，因此有助于T2V任务的视觉质量。作为NUWA的默认设置，所有三项任务的训练都能获得最佳性能。

标签。7显示了VSPW[24]数据集上草图到视频（S2V）任务的3D邻近注意的有效性。因为我们在研究S2V视频编码器和解码器时，都使用了3D数据。为了评估S2V的语义一致性，我们提出了一种称为检测到的PA的新指标，该指标使用语义分段模型[49]]分割生成视频的每一帧，然后计算生成片段和输入视频草图之间的像素精度。最后一行的默认NUWA设置，包括附近的编码器和附近的解码器，可实现最佳FID vid和检测到的PA。如果编码器或解码器被全神贯注替换，则性能会下降，这表明关注附近的条件和附近生成的结果比简单地考虑所有信息要好。我们将相邻稀疏和轴向稀疏进行了两次比较。首先，计算复杂度的附近稀疏和轴稀疏注意*O*（（硬件）（h+w+s））。对于生成长视频（较大），邻近的稀疏将更高效。其次，邻近稀疏在视觉生成任务中比轴稀疏具有更好的性能，这是因为附近稀疏参与了空间和时间轴之间的相互作用的“附近”位置，而轴稀疏单独处理不同的轴，并且只考虑在同一轴上的相互作用。*s*

图9显示了本文提出的一项新任务，我们称之为“文本引导视频操作（TV2V）”。TV2V的目标是从文本引导的选定帧开始改变视频的未来。所有样本从第二帧开始改变视频的未来。第一行显示原始视频帧，其中潜水员正在水中游泳

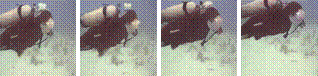
表6。VSPW数据集上草图到视频（S2V）任务的3D邻近注意的有效性。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 编码器 | 译码器 | 氢火焰离子化检测器↓ | 检测到PA↑ |
| 女娲-FF¨ | 满的 | 满的 | 35.21 | 0.5220 |
| 女娲岛 | 附近的 | 满的 | 33.63 | 0.5357 |
| 女娲-FN¨ | 满的 | 附近的 | 32.06 | 0.5438 |
| 女娲 | 轴 | 轴 | 29.18 | 0.5957 |
| 女娲 | 附近的 | 附近的 | 27.79 | 0.6085 |

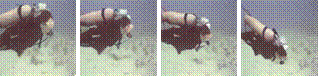
原始视频：



操作1：潜水员正在游到水面。



操作2：潜水员正在游到底部。



操纵3：潜水员飞向天空

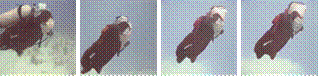


图9。同一视频上不同操作的示例。

水将“潜水员游泳到水面”输入NUWA的编码器并提供第一个视频帧后，NUWA成功生成潜水员游泳到第二排水面的视频。第三排显示了另一个成功的样本，让潜水员游到底。如果我们想让潜水员飞上天空呢？第四排显示女娲也能成功，潜水员像火箭一样向上飞行。

# 5.结论

在本文中，我们提出了NUWA作为一个统一的预训练模型，可以生成新的或操作现有的图像和视频，用于8个视觉合成任务。本文做出了一些贡献，包括（1）一个通用的3D编码器框架，同时涵盖文本、图像和视频；（2） 邻近稀疏注意机制，同时考虑空间轴和时间轴的邻近特性；（3） 8个综合任务的综合实验。这是我们朝着构建一个AI平台迈出的第一步，该平台能够创建视觉世界并帮助内容创作者。

# 工具书类

[1] 穆罕默德·巴贝伊扎德、切尔西·芬恩、杜米特鲁·埃尔汉、罗伊·H·坎贝尔和谢尔盖·莱文。随机变分视频预测。arXiv预印本arXiv:1710.112522017。6.

[2] 约格什·巴拉吉、马丁·人民、白冰、拉玛·切拉帕和汉斯·彼得·格拉夫。用于文本到视频合成的带鉴别滤波器生成的条件GAN。IJCAI，第1995-2001页，2019页。6.

[3] 大卫·包、亚历克斯·安多尼亚、奥黛丽·崔、延焕公园、阿里·贾哈尼安、奥德·奥利瓦和安东尼奥·托拉巴。用文字作画。ARXIV预告ARXIV：2103.10951，2021。7.

[4] 安德鲁·布鲁克、杰夫·多纳休和凯伦·西蒙扬。用于高保真自然图像合成的大规模GAN训练。arXiv:1809.11096[cs，统计]，2019年2月。1.

[5] Mark Chen、Alec Radford、Rewon Child、Jeffrey Wu、Heewoo Jun、David Luan和Ilya Sutskever。从像素生成预训练。在国际机器学习会议上，第1691-1703页。PMLR，2020年。1, 2

[6] 雷翁·查尔德、斯科特·格雷、亚历克·雷德福德和伊利亚·萨茨基。使用稀疏变换器生成长序列。arXiv预印本arXiv:1904.105092019。2.

[7] 凯文·克拉克、乌尔瓦希·坎德尔瓦尔、奥马尔·利维和克里斯托弗·曼宁。伯特在看什么？对伯特注意力的分析。arXiv预印本arXiv:1906.04341199。6.

[8] 艾米丽·丹顿和罗布·弗格斯。具有学习先验知识的随机视频生成。在国际机器学习会议上，第1174-1183页。PMLR，2018年。6.

[9] 丁明、杨卓义、洪文义、郑文迪、常州、大音、林俊阳、徐邹、周绍、杨红霞、唐杰。CogView：通过变压器控制文本生成图像。ARXIV：2105.13290 [C]，2021年5月。2, 6, 12

[10] 弗雷德里克·埃伯特、切尔西·芬恩、亚历克斯·李和谢尔盖·莱文。具有时间跳跃连接的自监督视觉规划。在CoRL，第344-356页，2017年。6.

[11] 帕特里克·埃塞尔、罗宾·罗姆巴赫和比约恩·奥默。¨用于高分辨率图像合成的驯服变压器。ARXIV：2012.09841 [C]，2021年6月。2, 3, 6, 12

[12] 切尔西·芬恩、伊恩·古德费罗和谢尔盖·莱文。通过视频预测进行物理交互的无监督学习。神经信息处理系统的进展，29:64–722016。6.

[13] Jean-Yves Franceschi、Edouard Delasales、Mickeel Chen、Sylvain Lamprier和Patrick Gallinari。随机潜在残差视频预测。国际机器学习会议，第3233-3246页。PMLR，2020年。6.

[14] 马丁·赫塞尔、休伯特·拉姆绍尔、托马斯·恩特丁纳、伯恩哈德·奈斯勒和塞普·霍克雷特。由两时间尺度更新规则训练的GAN收敛到局部纳什均衡。神经信息处理系统的进展，2017年3月30日。7.

[15] Jonathan Ho、Nal Kalchbrenner、Dirk Weissenborn和Tim Salimans。多维变压器的轴向注意。arXiv预印本arXiv:1912.121802019。2, 11

[16] 威尔·凯、乔·卡雷拉、卡伦·西蒙扬、布莱恩·张、克洛伊·希利尔、苏德赫德拉·维贾亚纳拉西姆汉、法比奥·维奥拉、，

蒂姆·格林、特雷弗·巴克和保罗·纳采夫。人体动作视频数据集。arXiv预印本arXiv:1705.06950，2017年。6.

[17] 迪德里克·金马和吉米·巴。Adam：随机优化的一种方法。arXiv预印本arXiv:1412.69802014。4.

[18] 马诺伊·库马尔、穆罕默德·巴贝伊扎德、杜米特鲁·埃尔汉、切尔西·芬恩、谢尔盖·莱文、劳伦特·丁赫和杜克·金马。Videoflow：一种基于条件流的随机视频生成模型。arXiv预印本arXiv:1903.014342019。6.

[19] 阿丽娜·库兹涅佐娃、哈桑·罗姆、尼尔·奥尔德林、贾斯珀·尤伊林斯、伊万·克拉辛、乔迪·庞特·塔塞特、沙哈布·卡马里、斯特凡·波波夫、马特奥·马洛西和亚历山大·科列斯尼科夫。打开图像数据集v4。国际计算机视觉杂志，128（7）：1956-1981，2020。7.

[20] Alex X.Lee、Richard Zhang、Frederik Ebert、Pieter Abbeel、Chelsea Finn和Sergey Levine。随机对抗视频预测。arXiv预印本arXiv:1804.01523，2018年。6.

[21]李易通、马丁·敏、沈丁汉、大卫·卡尔森和劳伦斯·卡林。从文本生成视频。《AAAI人工智能会议记录》，第32卷，2018年。6.

[22]Tsung Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollar和C.Lawrence&apos;Zitnick。Microsoft coco：上下文中的常见对象。欧洲计算机视觉会议，第740-755页。斯普林格，2014年。4, 5, 6, 7

[23]Pauline Luc、Aidan Clark、Sander Dieleman、Diego de Las Casas、Yotam Doron、Albin Cassier和Karen Simonyan。基于变换的大规模数据对抗性视频预测。arXiv预印本arXiv:2003.040352020。6.

[24]苗家旭、魏云超、于武、陈亮、李光瑞、杨毅。VSPW：用于野外视频场景解析的大规模数据集。在IEEE/CVF计算机视觉和模式识别会议上，第4133页至第4143, 2021页。7, 8

[25]纪尧姆·勒莫、让·庞塞和科迪利亚·施密德。CCVS：上下文感知可控视频合成。ARXIV预告ARXIV：2107.08037，2021。6.

[26]Mathew Monfort、Alex Andonian、Bolei Zhou、Kandan Ramakrishnan、Sarah Adel Bargal、Tom Yan、Lisa Brown、Quanfu Fan、Dan Gutfreund和Carl Vondrick。时刻数据集：一百万个视频用于事件理解。IEEE模式分析和机器智能学报，42（2）：502-50812019。4.

[27]朴大成、刘明宇、王廷春和朱俊彦。具有空间自适应归一化的语义图像合成。IEEE/CVF计算机视觉和模式识别会议记录，第2337-23462019页。6.

[28]Niki Parmar、Ashish Vaswani、Jakob Uszkoriet、Lukasz Kaiser、Noam Shazeer、Alexander Ku和Dustin Tran。图像转换器。arXiv预印本arXiv:1802.057511918。1.

2

[29]亚历克·拉德福德、金钟郁、克里斯·哈拉西、阿迪蒂亚·拉梅什、加布里埃尔·吴、桑迪尼·阿加瓦尔、吉里什·萨斯特里、阿曼达·阿斯克尔、帕梅拉·米什金、杰克·克拉克、格雷琴·克鲁格和伊利亚·萨茨基佛。从自然语言监督中学习可转移的视觉模型。ARXIV：2103.00020 [C]，2月2021日。5.

[30]亚历克·拉德福德、卢克·梅茨和苏米特·钦塔拉。深度卷积生成对抗网络的无监督表征学习。arXiv预印本arXiv:1511.064342015。1.

[31]Ruslan Rakhimov、Denis Volkhonskiy、Alexey Artemov、Denis Zorin和Evgeny Burnaev。潜在的视频变压器。arXiv预印本arXiv:2006.107042020。2, 6, 11

[32]普拉吉特·拉马钱德兰、尼基·帕尔玛、阿什什·瓦斯瓦尼、欧文·贝洛、安瑟姆·列夫斯卡娅和乔纳森·什伦斯。视觉模型中的独立自我注意。arXiv预印本arXiv:1906.05909，2019年。2.

[33]阿迪蒂亚·拉梅什、米哈伊尔·巴甫洛夫、加布里埃尔·高、斯科特·格雷、切尔西·沃斯、亚历克·拉德福德、马克·陈和伊利亚·萨茨克尔。零镜头文本到图像生成。ARXIV：2102.12092 [C]，2月2021日。1, 2, 5, 6, 11

[34]奥尔加·鲁萨科夫斯基、贾登、郝苏、乔纳森·克劳斯、桑吉夫·萨提什、肖恩·马、黄志恒、安德烈·卡帕蒂、阿迪蒂亚·科斯拉、迈克尔·伯恩斯坦、亚历山大·贝格和李菲菲。ImageNet大规模视觉识别挑战。arXiv:1409.0575[cs]，2015年1月。7.

〔35〕Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford、陈曦。改进了训练gans的技术。神经信息处理系统的进展，29:2234–224220016。5.

[36]陶明、唐昊、吴松松、倪思柏、肖元静、吴飞、包秉坤。Df-gan：用于文本到图像合成的深度融合生成对抗网络。arXiv预印本arXiv:2008.058652020。6.

[37]Sergey Tulyakov、刘明宇、杨晓东和Jan Kautz。Mocogan：分解运动和内容以生成视频。《IEEE计算机视觉和模式识别会议记录》，第1526-1535页，2018年。1, 6

[38]托马斯·恩特丁纳、舍尔德·范·斯滕基斯特、卡罗尔·库拉赫、拉斐尔·马里尼尔、马辛·米查尔斯基和西尔万·盖利。视频的精确生成模型：一个新的度量和挑战。arXiv预印本arXiv:1812.01717，2018年。6.

[39]亚伦·范登奥尔德、纳尔·卡尔希布伦纳、奥利奥·维尼亚、拉塞·埃斯佩霍尔特、亚历克斯·格雷夫斯和科雷·卡武科库奥卢。使用像素CNN解码器生成条件图像。arXiv预印本arXiv:1606.053281016。1, 2

[40]Aaron van den Oord、Oriol Vinyals和Koray

Kavukcuoglu。神经离散表示学习。arXiv预印本arXiv:1711.009372017。1, 2

[41]Aaron Van Oord、Nal Kalchbrenner和Koray Kavukcuoglu。像素递归神经网络。在国际机器学习会议上，第1747-1756页。PMLR，2016年。1, 2

[42]Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkorit、Llion Jones、Aidan N.Gomez、Lukasz Kaiser和Illia Polosukhin。注意力是你所需要的。《神经信息处理系统的进展》，第5998–6008页，2017年。2.

[43]王欣、吴家伟、陈俊坤、李磊、王元芳和王威廉·杨。Vatex：用于视频和语言研究的大规模、高质量的多语言数据集。《IEEE/CVF计算机视觉国际会议记录》，第4581-45912019页。4.

[44]德克·魏森伯恩、奥斯卡·塔克斯特罗姆和雅各布·乌兹科雷特。¨缩放自回归视频模型。在ICLR，2020年。1, 2, 6, 11

[45]吴晨飞、黄伦、张黔西、李宾阳、雷吉、范扬、吉列尔莫·萨皮罗、南段。GODIVA：从自然描述生成开放域视频。ARXIV：2104.14806 [CS]，4月2021日。1, 2, 5, 11

[46]徐军、陶梅、姚婷和永瑞。Msr vtt：用于连接视频和语言的大型视频描述数据集。《IEEE计算机视觉和模式识别会议记录》，第5288-52962016页。8.

[47]徐涛、张鹏川、黄秋元、张汉、詹哲、黄晓雷、何晓东。Attngan：具有注意力生成对抗网络的细粒度文本到图像生成。《IEEE计算机视觉和模式识别会议记录》，第1316-13242018页。1, 6

[48]颜伟信、张云芝、彼得·阿比尔和阿拉文·斯里尼瓦斯。VideoGPT：使用VQ-VAE和

变形金刚。ARXIV预告ARXIV：2104.10157，2021。2, 3

[49]袁玉辉、陈锡林、王景东。用于语义分割的对象上下文表示。计算机视觉-ECCV 2020：第16届欧洲会议，英国格拉斯哥，2020年8月23日至28日，会议记录，第六部分，第16页，第173-190页。斯普林格，2020年。8.

[50]张汉、郭靖宇、贾森·鲍德里奇、李洪拉克和杨银飞。跨模态对比学习用于文本图像生成。在IEEE/CVF计算机视觉和模式识别会议上，第833页至第842, 2021页。6.

[51]张云芝、颜伟信、彼得·阿比尔和阿拉文·斯里尼瓦斯。VideoGen：使用VQ-VAE和Transformers生成视频模型。2020年9月。3.

[52]朱民锋、潘平波、陈伟、杨毅。Dm-gan：用于文本图像合成的动态记忆生成对抗网络。IEEE/CVF计算机视觉和模式识别会议记录，第页

5802–5810, 2019. 6.

|  |
| --- |
| 图10。不同三维稀疏模型的比较。所有样本均假定输入3D数据的大小为4×4×2=32。上半部分的插图显示了生成目标令牌（橙色）需要使用哪些令牌（蓝色）。下半部分中大小为32×32的矩阵显示稀疏注意中的注意掩码（黑色表示掩码标记）。 |

表7。不同3D稀疏注意的复杂性。

|  |  |
| --- | --- |
| 单元 | 复杂性 |
| 三维完整  三维块稀疏[31,44]  三维轴向稀疏[15,33,45]  3D近距离稀疏（我们的） |  |

# A.三维稀疏图像之间的比较

图10显示了不同3D稀疏注意之间的比较。假设我们有大小为4×4×2的3D数据，3D块稀疏注意的思想是将3D数据分割成几个固定的块，并分别处理这些块。分割块的方法有很多，例如在时间、空间或两者中分割。图10中的3D块稀疏示例考虑时间和空间的分割。三维数据分为4个部分，每个部分的大小为2×2×2。要生成橙色标记，3D块稀疏注意将考虑固定3D块内以前的标记。尽管3D块稀疏注意同时考虑了空间轴和时间轴，但是这种空间和时间信息在3D块中是有限和固定的，特别是对于沿着3D块边缘的令牌。仅考虑部分邻近信息，因为3D块外部的某些邻近信息对于其中的令牌是不可见的。3D轴向稀疏注意力的想法是考虑沿着轴线的先前标记。尽管3D轴稀疏注意同时考虑了空间轴和时间轴，但这种空间和时间信息沿轴是有限的。仅考虑部分附近信息，而不在轴中的一些附近信息将不会在3D轴注意中考虑。在本文中，我们提出了一种三维邻近稀疏算法，该算法考虑了完整的邻近信息，并为每个令牌动态生成三维邻近注意块。注意矩阵也显示了证据，因为3D邻近稀疏的关注部分（蓝色）比3D区块稀疏和3D轴向稀疏更平滑。

标签。7显示了不同3D稀疏注意的复杂性。表示三维数据的空间高度、空间宽度和时间长度。不同的稀疏机制在不同的场景下有其计算优势。例如，对于长视频或具有较大分辨率的高分辨率帧*h、 w，sh、 w，s*通常，三维邻近稀疏注意比三维轴向稀疏注意更有效。如果三维数据可以拆分为多个部分而没有依赖关系，那么三维块稀疏将是一个不错的选择。例如，一部卡通有几集，每集讲述一个独立的故事，我们可以简单地分割这些故事，因为它们没有任何关系。

表8。NUWA的两个设置的实现细节。¨

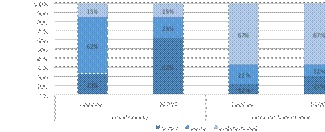
|  |  |  |
| --- | --- | --- |
| 设置 | 女娲-256¨ | 女娲-336¨ |
| VQGAN图像分辨率 | 256×256 | 336×336 |
| VQGAN离散令牌 | 32×32 | 21×21 |
| VQGAN压缩比 | F8 | F16 |
| VQGAN码本维数 | 256 | 256 |
| 3DNA隐藏大小 | 1280 | 1280 |
| 3DNA头数 | 20 | 20 |
| 每个头部的3DNA尺寸 | 64 | 64 |
| NUWA编码器层– | 12 | 12 |
| 女娲解码器层– | 24 | 24 |
| 多任务预训练数据集 | 概念说明  时间瞬间Vatex | |
| 多任务输入三维尺寸 | 1×1×77（T2I）  32×32×1（V2V） | 1×1×77（T2I）  21×21×1（V2V） |
|  | 1×1×77（T2V） | 1×1×77（T2V） |
| 多任务输出三维尺寸 | 32×32×1（T2I）  32×32×4（V2V） | 21×21×1（T2I）  21×21×10（V2V） |
|  | 32×32×4（T2V） | 21×21×10（T2V） |
| 训练批量 | 128 | |
| 培训学习率 | 10−3. | |
| 训练步骤 | 50米 | |

# B.多任务预培训的详细信息

标签。8显示了本文中使用的两个NUWA设置的实现细节。NUWA-256和N¨UWA-¨336型号都在图像质量和视频长度（视频帧数）之间进行权衡。由于图像质量高度依赖于压缩比和离散标记的数量，低压缩比和大离散标记是高质量图像的关键因素。然而，由于模型的总容量有限，每个图像的离散令牌数量和视频帧（图像）数量是折衷的。

请注意，NUWA-256采用F8压缩比，离散令牌为32×32，而NUWA-336采用F8压缩比，离散令牌仅为21×21。为了与当前最先进的模型进行公平比较，我们采用了NUWA-256和更离散的标记来生成高质量的图像。但是，考虑到transformer的效率，NUWA-¨256只能生成4帧的视频。要处理相对较长的视频，NUWA-336使用较少的离散标记可以生成10帧的视频。因此，NUWA-336在第二阶段显著降低了自回归模型的压力，尤其是视频。NUWA-336是覆盖图像和视频的默认设置。

对于这两种模型，请注意，我们没有过度调整参数，只是使用相同的学习率10−3和50米训练步骤。



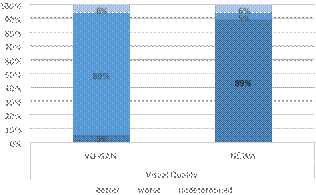
图11。针对文本图像（T2I）任务，对MSCOCO数据集进行人工评估。

图12。MSCOCO图像数据集的人体评价

完成（I2I）任务。

# C.人的评价

图11显示了CogView[9]和我们的NUWA在MSCOCO数据集–用于文本到图像（T2I）任务上的人类比较结果。我们随机选择了2000个文本，并要求注释者比较两个模型的生成结果，包括视觉质量和语义一致性。注释者被要求在三个选项中进行选择：更好、更差或待定。在视觉质量部分，62%的人支持我们的NUWA？模型，15%的人不确定，23%的人支持CogView，这表明NUWA生成了更真实的图像。在语义一致性部分，尽管67%的选票无法确定哪种模式更符合文本，女娲也赢得了剩余的21%的选票。¨尽管CogView在比NUWA更大的文本-图像对上进行了预训练，但我们的模型仍然受益于多任务预训练，因为文本-视频对为文本-图像生成提供了高级语义信息。

图12显示了VQGAN[11]和我们在MSCOCO数据集上的NUWA模型之间的人为比较结果，用于图像完成（I2I）任务。我们使用与图11类似的设置，但删除了语义一致性，因为此任务没有文本输入。比较结果显示，女娲获得89%的选票，显示出女娲强大的零射能力。¨

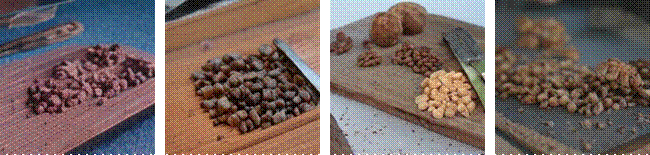
坐落在田野里的木屋。



一个年轻女孩正在吃一块看起来很美味的比萨饼。



核桃正在木制砧板上切割。



戴帽子戴领带的男孩。

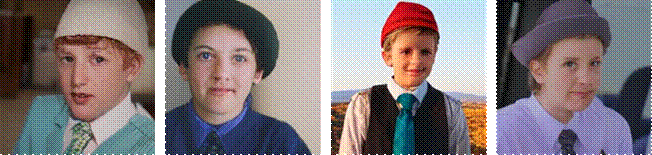
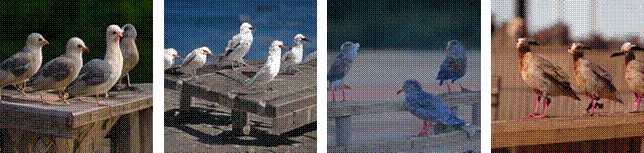


图13。NUWA生成的文本到图像（T2I）任务的更多示例。¨

一些鸟站在木凳上。



一只戴着圣诞老人帽子的狗躺在床上。



一个孩子正坐在一块插着蜡烛的蛋糕前。



一碗有肉酱、花椰菜和黄瓜的食物。

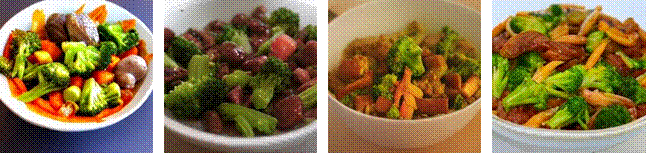


图14。NUWA生成的文本到图像（T2I）任务的更多示例。¨

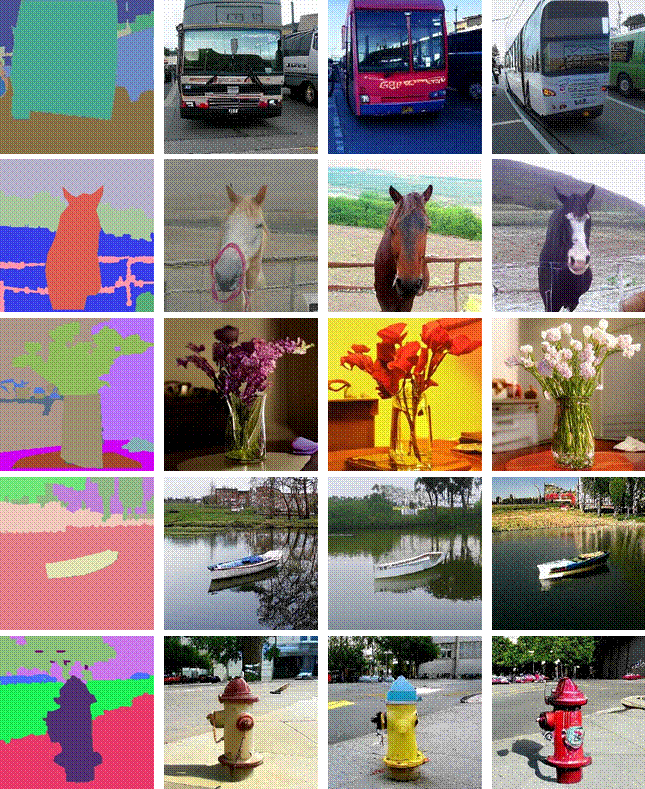
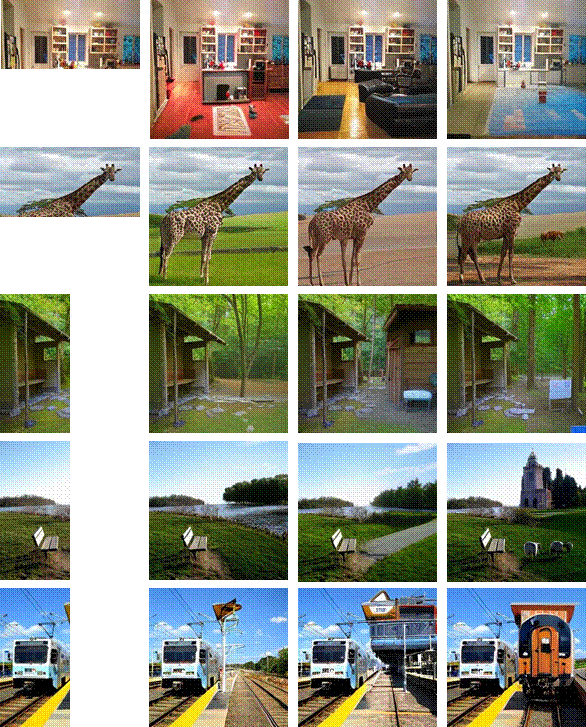


图15。NUWA生成的草图到图像（S2I）任务的更多示例。¨



图16。NUWA生成的草图到图像（S2I）任务的更多示例。¨

50%遮罩

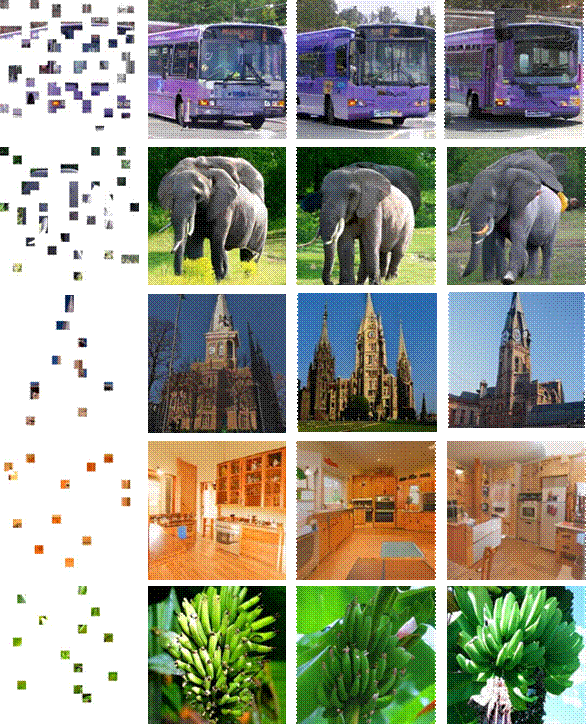
50%遮罩

50%遮罩

50%遮罩

50%遮罩

图17。NUWA生成的图像完成（I2I）任务的更多示例。¨

75%遮罩

85%遮罩

95%面罩

95%面罩

95%面罩

图18。NUWA生成的图像完成（I2I）任务的更多示例。¨

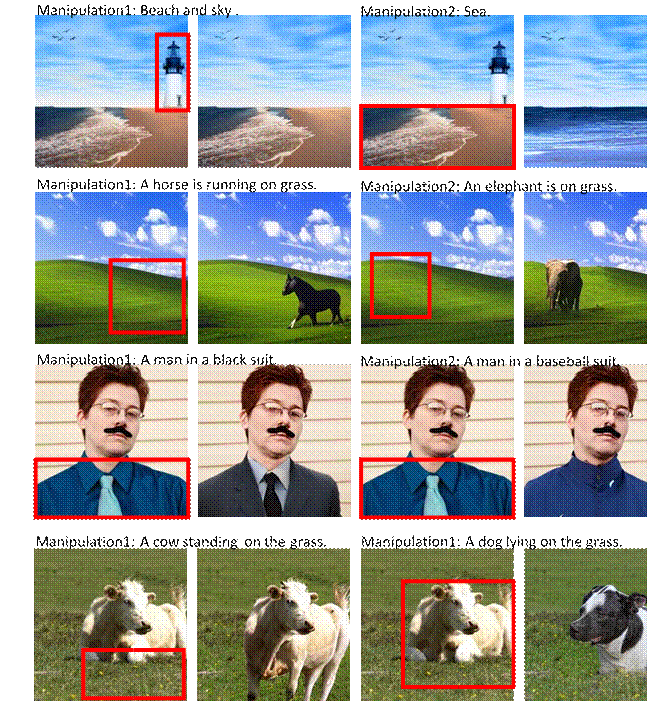


图19。NUWA生成的文本引导图像处理（TI2I）任务的更多示例。¨

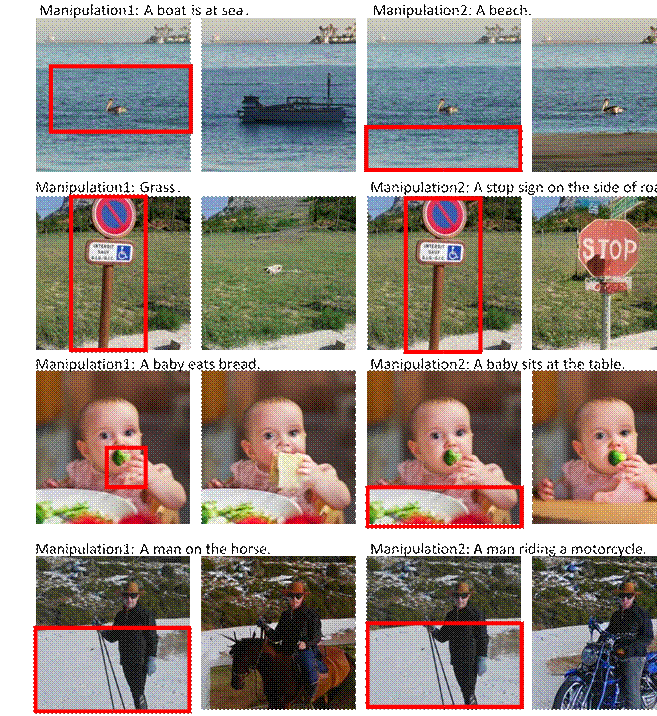


图20。NUWA生成的文本引导图像处理（TI2I）任务的更多示例。¨

在草地上打高尔夫球。



在游泳池打高尔夫球。



在海上奔跑。

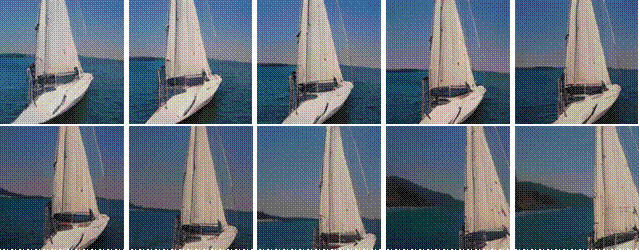


图21。NUWA生成的文本到视频（T2V）任务的更多示例。¨

一个穿西装的人在演播室里谈笑风生。



白色的帆船在海上航行。



一个人正在折一张黄色的纸。

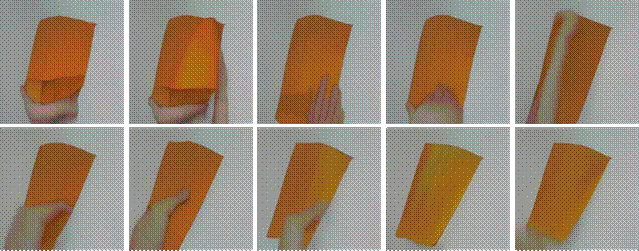


图22。NUWA生成的文本到视频（T2V）任务的更多示例。¨

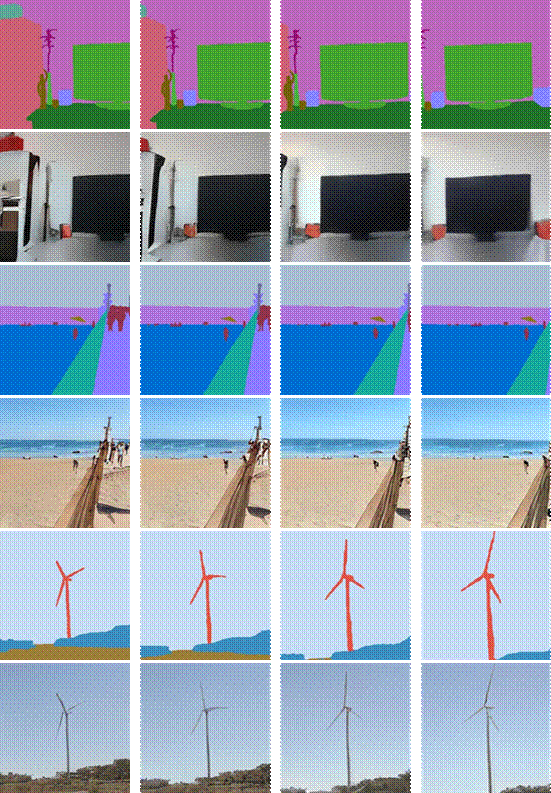


图23。NUWA生成的更多草图到视频（S2V）任务示例。¨

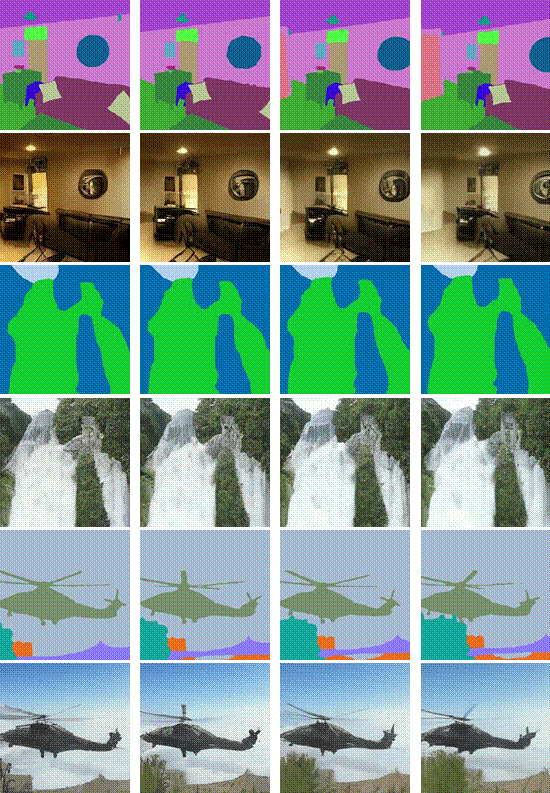


图24。NUWA生成的更多草图到视频（S2V）任务示例。¨

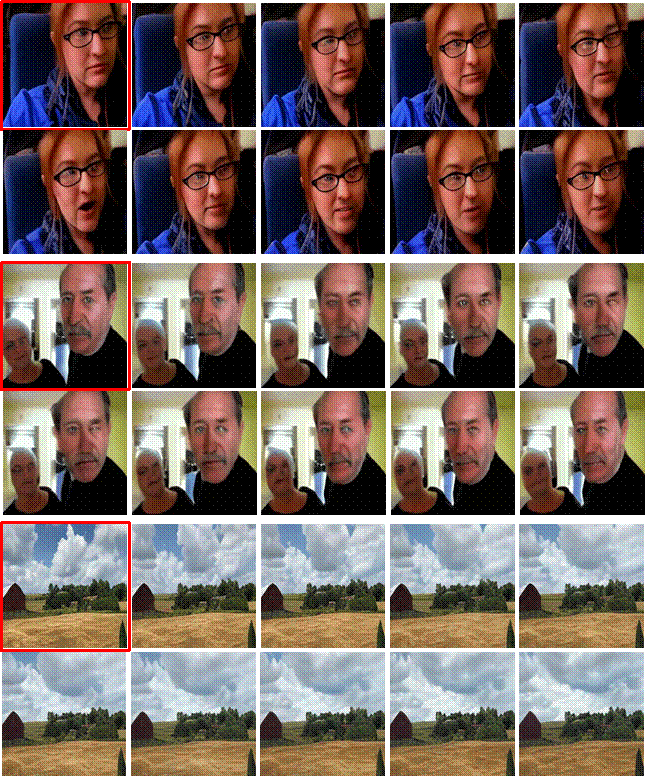


图25。NUWA生成的视频预测（V2V）任务的更多示例。只有一个帧（见红色框）用作条件。¨

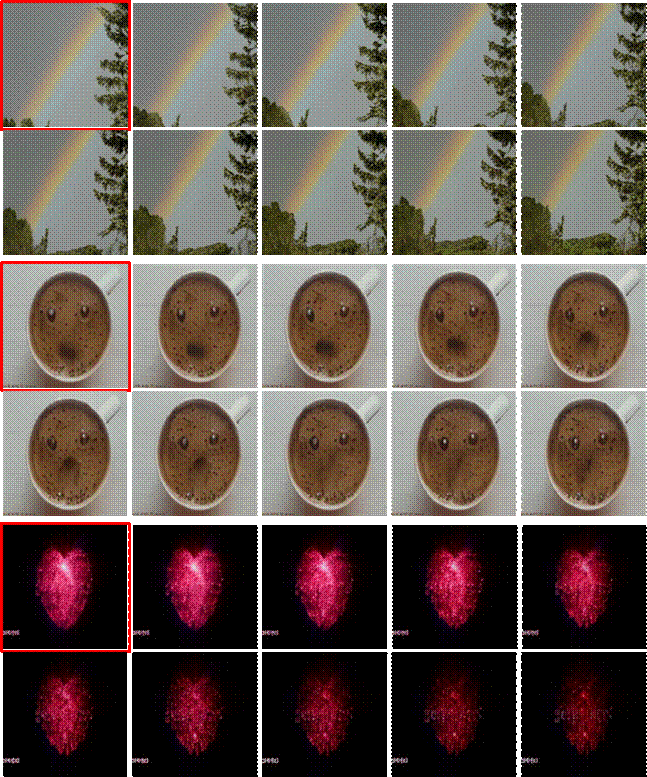
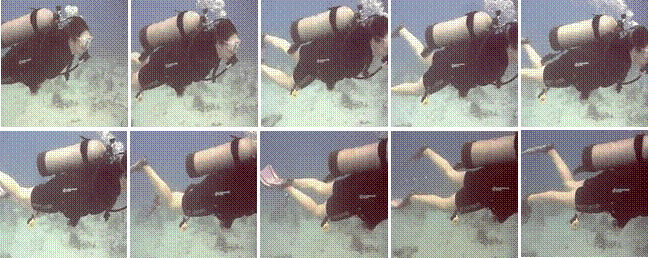


图26。NUWA生成的视频预测（V2V）任务的更多示例。只有一个帧（见红色框）用作条件。¨

原始视频：



操作1：潜水员正在游到水面。

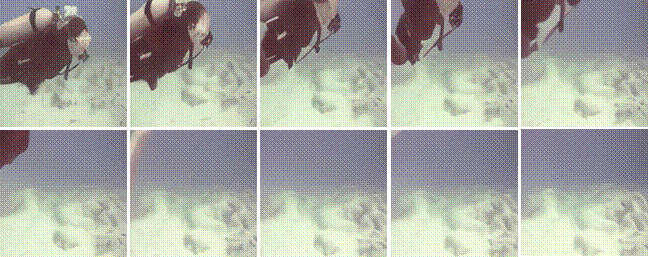
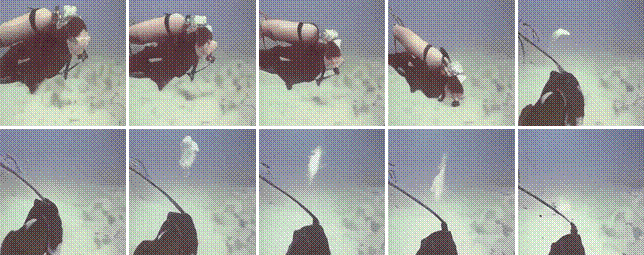


图27。NUWA生成的更多文本引导视频操作（TV2V）任务示例。¨

操作2：潜水员正在游到底部。



操纵3：潜水员正飞向天空。



图28。NUWA生成的更多文本引导视频操作（TV2V）任务示例。¨