

实例扩散模型的研究报告

摘要

文本到图像的扩散模型能生成高质量的图像，但无法控制图像中的单个实例。通过引入实例扩散（InstanceDiffusion）为文本到图像扩散模型增加了精确的实例级控制。InstanceDiffusion 支持每个实例的自由格式语言条件，并允许以灵活的方式指定实例负载，如简单的单点、涂鸦、边界框或复杂的实例分割掩码，以及它们的组合。并且针对传统的文本-图像模型提出了三大变革，以实现精确的实例级控制。其中，UniFusion 模块实现了文本到图像模型的实例级条件，ScaleU 模块提高了图像的清晰度，多实例采样器改进了多实例的生成。InstanceDiffusion 在每个位置条件下都大大超越了这个领域的 sota。具体来说，在 COCO 数据集上，对于边界框输入，InstanceDiffusion 的表现比之前的 sota 在指标 AP_{box}^{50} 高出 20.4%，对于掩码输入，比之前的 sota 的 IoU 提升了 25.4%。

关键词：实例级扩散；图像生成

1 引言

近年来，基于网络规模数据训练的图像生成模型，如文献 [5, 6, 15, 17, 19, 20, 31, 32, 34, 37, 47]，取得了显著进展。这些工作在 GAN、Diffusion 或是 Transformer 的基础上做出改进，并在零样本 COCO 数据集上的验证，指标如 FID 等随着工作的迭代而更新。在这些工作的基础之上，不少人还在探索如何根据给定条件生成对应更贴合人类想法的图像。例如，基于文本条件扩散模型现在可以生成包含文本中指定的自由形式概念的高质量图像 [9, 17, 31, 34, 37, 38]。虽然基于文本的控制能够生成质量更好的图像，但它并不总能对输出图像进行精确和直观的控制。因此，为了更好地控制，人们提出了许多不同形式的调节方法，如边缘、法线图、语义布局等 [2, 4, 11, 12, 14, 24, 28, 29, 43, 44]。这些更丰富的控制使生成模型在设计、数据生成 [13, 46] 等方面的应用更加广泛。本项工作 InstanceDiffusion 的重点在于精确控制实例在输出图像中的位置和属性。提出并研究了以实例为条件的图像生成方法，用户可以根据实例的位置和实例级文本提示来指定每个实例，从而生成图像。可以使用边界框、实例掩码、单点或涂鸦来指定位置。这种输入方式非常灵活，有些实例的位置可以使用掩码来指定得更精确，而有些则可以使用点来指定得不那么精确。每个实例的文本提示允许对实例的属性（如颜色、纹理等）进行精细控制。本文所提出的实例条件生成是对先前工作 [2, 24, 44] 中研究的设置的概括，这些工作只考虑一种位置格式，不使用每个实例的标签。本文模型提供了若干种位置格式，包括简单的单点、边界框、掩码以及涂鸦，可以更精确、更灵活地控制输出图像中的实例。为此提出了一种统一位置格式的方法，即在生成过程中对位置信息进行参数化和融合。与之前使用不同架构和策略对不同位置格式进行建模的方式相比，统一建模显然使得模型结构变得

更简单。此外，位置格式的统一建模允许模型利用实例位置的共享底层结构，从而提高性能。通过综合评估，InstanceDiffusion 优于专门针对特定实例条件的 sota。在对 COCO [25] val 上的边界框输入进行评估时，本文模型的 AP_{50}^{box} 比 GLIGEN [24] 提高了 20.4%。对于基于掩码的输入，IoU 比 DenseDiffusion [21] 提高了 25.4%， AP_{50}^{mask} 比 ControlNet [44] 提高了 36.2%。由于之前的方法没有研究用于图像生成的点或涂鸦输入，因此引入了针对这些设置的评估指标。InstanceDiffusion 在遵循实例级文本提示指定的属性方面也表现出了卓越的能力。与 GLIGEN 相比，在实例颜色准确性方面提高了 25.2%，在纹理准确性方面提高了 9.2%。本文的主要贡献如下：

- (1) 提出并研究了以实例为条件生成图像的方法，这种方法允许灵活地指定多个实例的位置和属性。
- (2) 提出了三种可改善结果的关键模块：(i) UniFusion，它将各种形式的实例级条件投射到同一特征空间，并将实例级布局和描述注入视觉标记；(ii) ScaleU，重新校准 UNet 跳转连接特征中的主要特征和低频成分，增强模型精确遵循指定布局条件的能力；(iii) 多实例采样器，减少信息泄漏和多个实例（文本 + 布局）条件之间的混淆。
- (3) 使用预训练模型生成的实例级标题数据集，以及一套新的评估基准和指标，用于衡量基于位置的图像生成性能。
- (4) 与之前的工作相比，对不同位置格式的统一建模大大改进了结果。同时该研究结果还可以应用于以前的方法来提高它们的性能。

2 相关工作

2.1 图像扩散模型

图像扩散模型是今年图像生成领域的重要方法之一，这类模型通过逐步消除噪声或者引导反向扩散过程，从噪声中生成高质量的图像。[17, 36] 中通过从初始随机噪音图开始的迭代去噪步骤，学习文本到图像的生成过程。潜在扩散模型（LDM）[33, 40] 在变异自动编码器 [22, 40] 的潜在空间中执行扩散过程，以提高计算效率，并将文本输入编码为来自预训练语言模型 [29] 的特征向量。DALL-E 2 [31] 使用 CLIP 的图像空间合成图像。相比之下，Imagen [34] 直接扩散像素，无需潜在图像。此外，它还证明了仅在文本语料库中训练的通用大型语言模型（如 T5 [30]）在为图像生成进行文本编码方面具有惊人的功效。

2.2 具有空间控制的图像生成

具有空间控制的图像生成方法一种有条件的图像合成任务，文献 [2, 11, 12, 16, 18, 24, 26, 39, 44] 通过引入空间条件控制来引导图像生成过程。其中 SpaText [2] 和 Make-a-Scene [12] 着重将空间信息和语义信息相结合，从而提高图像生成模型的准确性和一致性。GLIGEN [24] 通过将文本条件和具体的空间信息结合，从而实现开放世界的条件图像生成，并且能够在零样本情况下完成图像生成，还支持使用离散条件（如边界框）进行控制的图像生成。ControlNet [44] 在大规模预训练扩散模型中添加了更细粒度的空间控制，例如语义分割掩码，允许用户包含明确定义所需图像构图的附加图像。MultiDiffusion [3]、DenseDiffusion [21]、Attend-and-Excite [7]、ReCo [42]、StructureDiffusion [10]、Layout-Guidance [8] 和 BoxDiff [41] 等方法则在无需对预

训练的文本-图像模型进行微调的情况下，为扩散模型添加了位置控制。

2.3 讨论

ControlNet 和 GLIGEN 都在提升图像生成的空间可控性方面做出了重要的贡献，但是它们需要为每种可控输入类型训练单独的模型，这增加了系统的整体复杂性，并且无法有效捕获各种可控输入之间的交互。此外，ControlNet 仅关注空间条件，需要精确的几何控制。而 GLIGEN 使用对象类别作为文本提示，缺乏基于详细实例级别描述的模型训练。这不仅限制了用户的控制能力，还阻碍了模型有效利用实例描述的能力。

3 本文方法

3.1 本文方法概述

首先，先讨论该方法的问题定义。该方法的目标是通过关注每个实例的两个条件输入，即其位置和描述该实例的文字说明，来改进图像生成中的实例级控制能力。用数学语言来描述这一目标，即学习一个图像生成模型 $f(c_g, \{(c_1, l_1), \dots, (c_n, l_n)\})$ ，该模型以全局文本标签 c_g 和每个实例条件 (c_i, l_i) 作为条件，其中 c_i 是实例的文本标签， l_i 是实例的位置，共有 n 个实例。这个问题类似于 SpaText [2] 中，“开放集基础文本到图像” [24] 问题的泛化，允许用户提供更细粒度的描述。这种扩展提升了模型的灵活性，用户不仅可以控制场景中每个实例的属性，还能同时指定整体的场景布局。如图 1 所示，在 InstanceDiffusion 中，通过提供额外的实例级控制，增强了文本到图像模型。除了全局文本提示外，还允许在生成图像时指定成对的实例级提示及其位置。并且支持各种位置形式，从最简单的点、边界框和涂鸦到更复杂的掩码，以及它们的灵活组合。

InstanceDiffusion 是用于实例条件图像生成的扩散模型。由于获取大规模的（文本、图像）配对数据比获取（实例、图像）数据要容易得多，因此使用预先训练好的文本-图像 UNet 模型，并保持该模型的参数不变（将模型冻结）。在此基础上，引入了可学习的 UniFusion 块，以处理每个额外的实例条件输入。UniFusion 将实例调节与主干网络融合，并调节其特征以实现实例条件图像生成。此外，还提出了 ScaleU 模块，通过重新调整 UNet 中生成的跳跃连接和主干网络生成的特征图，提高了 UNet 遵循实例条件的能力。在推理过程中，采用多实例采样器（Multi-instance Sampler），可减少多个实例之间的信息泄露。由于难以获得大规模的（实例、图像）配对数据集，此项工作中利用最新的识别系统自动生成了一个包含实例级位置和文本标题的数据集。

最后，为了评估模型在实例条件生成任务中的表现提出了一个新的综合基准。

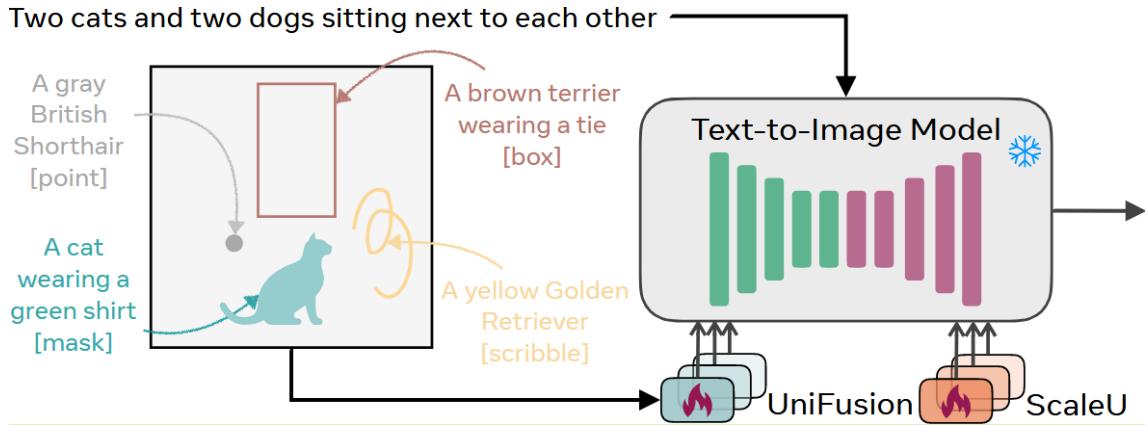


图 1. InstanceDiffusion

3.2 UniFusion 模块

如图 2 所示，UniFusion 模块将每个实例的条件 ($c_i l_i$) 转化为标记，并将其与来自冻结的文本-图像模型的特征（即视觉标记）融合。类似于 Flamingo [1] 和 Gligen [24], UniFusion 被插入到主干网络的自注意力层和交叉注意层之间。每个实例的位置 l_i 可以用一种或多种位置格式指定，例如掩码、边界框等。UniFusion 模块将完成以下重要操作。

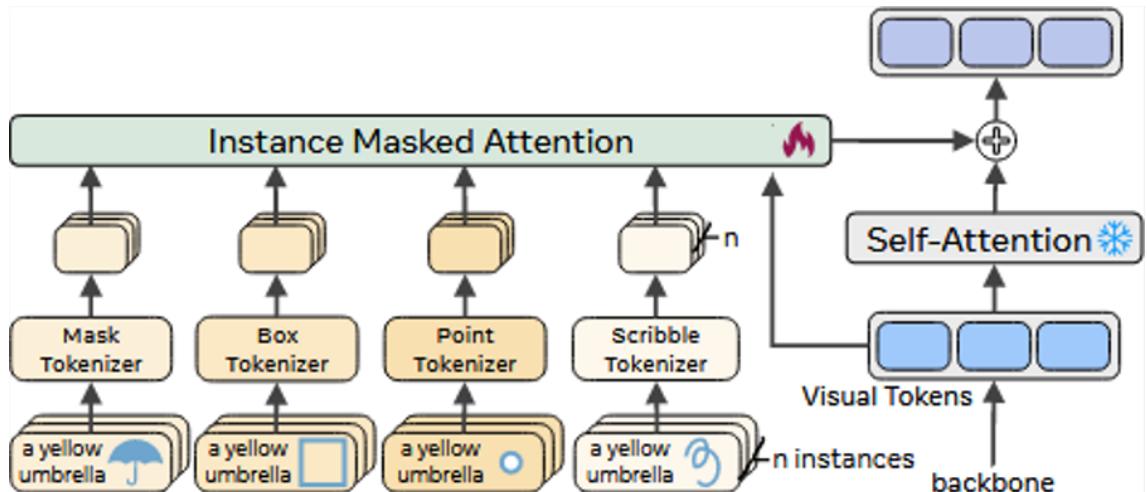


图 2. UniFusion

位置参数化。如图 3 所示，对于模型支持的四种位置格式，均转换为格式统一的二维点集。对于实例 i ，其位置格式可表示为 $p_i = (x_k, y_k)_{k=1}^n$ 。具体来说，涂鸦会被转换成一组沿曲线均匀采样的点；边界框以左上角和右下角进行参数化；将掩码转换为从掩码内和边界多边形中采样的点集；单点则无需转换操作。

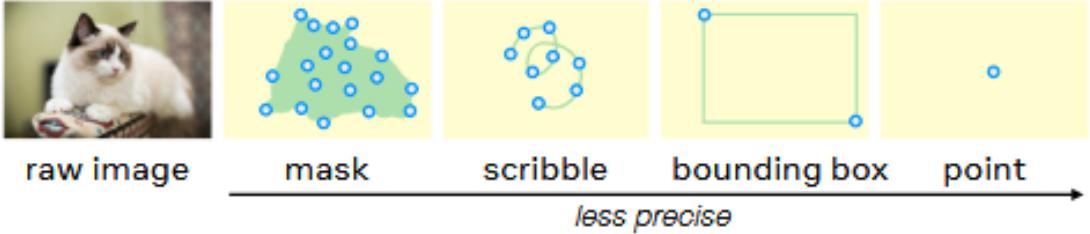


图 3. UniFusion 中的位置参数化

实例标记。通过位置参数化后所得的 p_i 使用傅里叶映射 $\gamma(\cdot)$ ，并通过 CLIP 文本编码器 $\tau_\theta(\cdot)$ 对文本提示 c_i 进行编码。最终将位置和文本嵌入进行拼接，并输入到 MLP，生成实例 i 的单个标记嵌入 $g_i = \text{MLP}([(c_i), (p_i)])$ 。值得注意的是，对于不同的位置格式应当使用不同的 MLP 网络。每个实例的位置 l_i 可以用一种或者多种位置格式，从而得到 g_i^{mask} 、 g_i^{scribble} 、 g_i^{box} 、 g_i^{point} 。

实例掩码注意力与融合机制。通过显式掩码限制不同实例间的交互，防止信息泄露。

$$\tilde{V} = SA_{\text{mask}}([V, G^{\text{mask}}, G^{\text{scribble}}, G^{\text{box}}, G^{\text{point}}])$$

其中， V 表示来自主干网络的 m 个视觉标记 v ， G 表示所有 n 个实例在每个位置格式下的条件标记 g 。这里采用联合格式，即将每种格式的嵌入拼接后通过一个 MLP 转换为单个嵌入，用于掩码自注意力，能够有效解决实例之间的信息泄露问题。

这一模块的设计克服了传统扩散模型在处理重叠实例和小目标上的不足，同时多种位置格式能够提供更加精细的条件控制，掩码自注意力机制有效解决了实例间信息泄露问题，从而有效提高生成结果的质量和准确性。

3.3 ScaleU 模块

在 Unet 模型中，每个块将主干特征图 F_b 和跳跃连接特征 F_s 合并后传递给后续的 Unet 块。FreeU [35] 揭示了扩散 UNet 尚待开发的潜力，能快速大幅提高生成质量。他们研究了 UNet 架构对去噪过程的主要贡献，发现其主干网主要对去噪做出贡献，而其跳过连接主要将高频特征引入解码器模块，导致网络忽略了主干网的语义。利用这一发现，提出了一种简单而有效的方法——FreeU，无需额外的训练或微调即可提高生成质量。通过减少跳跃特征中的低频成分，同时通过通道无关且经验调节的值增强主特征。由于 FreeU 仍无法根据实例条件动态调整特征的重要性。因此在 InstanceDiffusion 中，引入了 ScaleU 模块，其包含两个可学习的通道级缩放向量 s_b 和 s_s ，通过动态调整 F_b 和 F_s 的通道可学习向量，可以显著提升性能。并且 ScaleU 增加的参数量非常小（仅 0.01%），但性能提升非常明显。

对主干特征 F_b 的调整：

$$F'_b = F_b \otimes (\tanh(s_b) + 1)$$

即通过简单的通道级乘法缩放。

对跳跃连接特征 F_s 的调整：

$$F'_s = \text{IFFT}(\text{FFT}(F_s) \odot \alpha)$$

选择低频成分（小于 r_{thresh} ）在傅里叶域中缩放。其中 FFT 和 IFFT 分别是快速傅里叶变换和逆傅里叶变换， \odot 是逐元素乘法， α 是频率掩码。当 r 小于 r_{thresh} 时， $\alpha(r) = \tanh(s_s) + 1$ ，否则 $\alpha = 1$ 。

3.4 Multi-instance Sampler

使用多实例采样器是为了进一步减少多实例条件之间的信息泄露，提高生成图像质量。图4展示了多实例采样器的工作流程。首先对于每个实例，执行一次独立的去噪操作，持续 M 步（小于总体去噪步骤的 10%），以获取该实例的潜变量 L_I 。然后，将获取到的 n 个实例潜变量 $\{L_I^1, \dots, L_I^n\}$ ，与通过所有实例标记和文本提示生成的全局潜变量 L_G 做加权平均进行结合。最后对结合而成的潜变量作进一步的去噪操作。

通过先独立生成每个实例的潜变量得到局部信息，再与全部潜变量进行结合，这样就确保了局部细节和全局结构相协调。并且去噪的过程是分阶段进行，有效减少了在实例独立处理阶段的冲突，让生成的对象边界更加清晰，语义表达更加准确。同时又能结合全局信息来提升生成图像的质量。

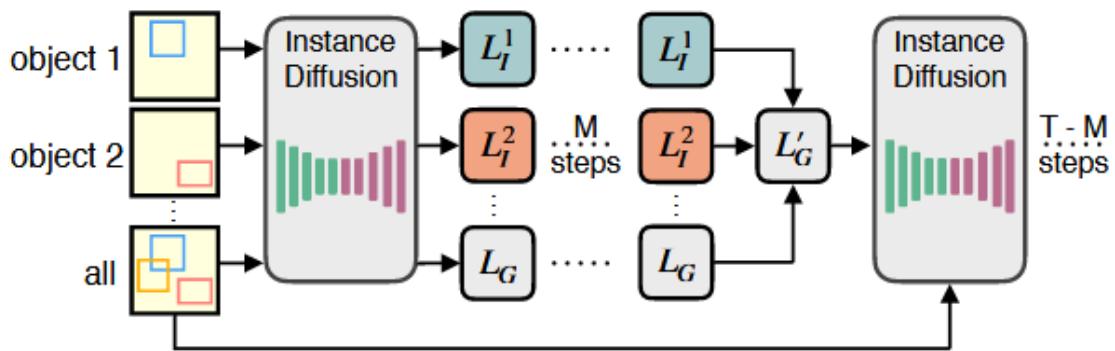


图 4. Multi-instance Sampler

4 复现细节

4.1 与已有开源代码对比

本工作基于 InstanceDiffusion 开源项目进行了完整的复现，在复现过程中采用了 `trainer.py`、`inference.py` 以及 `eval_local.py`。其中 `trainer.py` 用于模型的训练，`inference.py` 可利用预训练模型生成图像，`eval_local.py` 可用于测试模型性能。由于训练数据并未公开，而标准的物体检测数据集如 COCO 只包含了每个物体位置的稀疏类别标签，并不是本工作中需要的更具体的条件描述。为了解决这一问题，我采用了 RAM [45] 来生成图像级标签，Grounded-SAM [27] 将图像级标签细化成实例级，生成与这些标签对应的精确的边界框和掩码，并使用视觉语言模型 BLIP-V2 [23] 给裁剪后的实例生成详细的描述性文本，将实例标签转化成自然语言提示。这样就构造了一个具有多样性和细节表现力的数据集，为模型训练提供了高质量的训练数据。

我还利用训练好的模型进行生成了 5000 张图片，然后用 YOLOv8 进行评估，得到了 F1-Confidence 曲线，Precision-Recall 曲线，Confidence 曲线，混淆矩阵以及分割结果的可视化（如图 8）。记录了模型在验证集上的预测结果，包括每个目标的类别、边界框、掩码数据以及

置信度。

4.2 实验环境搭建

Ubuntu 20.04 with Cuda 12.4

python 3.12.2, pytorch 2.5.1, torchvision 0.20.1, torchaudio 2.5.1

4.3 界面分析与使用说明

操作界面如图 5 所示，需要通过输入命令行。如 Listing 1 所示，调用 inference.py 脚本，使用一个预训练模型权重和指定的输入文件来生成 8 张图像，并将生成的结果保存在对应文件夹中。命令中包含的其他参数（如 guidance_scale 和 alpha）用于控制生成图像的特性和条件权重。

```
1 # 训练模型后，生成图像。
2 python inference.py \
3     --num_images 8 \          #生成图像数量
4     --output OUTPUT/ \        #输出图像路径
5     --input_json INPUT/ \     #输入文本信息和文本格式
6     --ckpt instancediffusion_sd15.pth \ #使用模型的预训练权重
7     --guidance_scale 7.5 \    #指导尺度
8     --alpha 0.8             #混合比例参数
```

Listing 1: 生成图像 bash 命令行

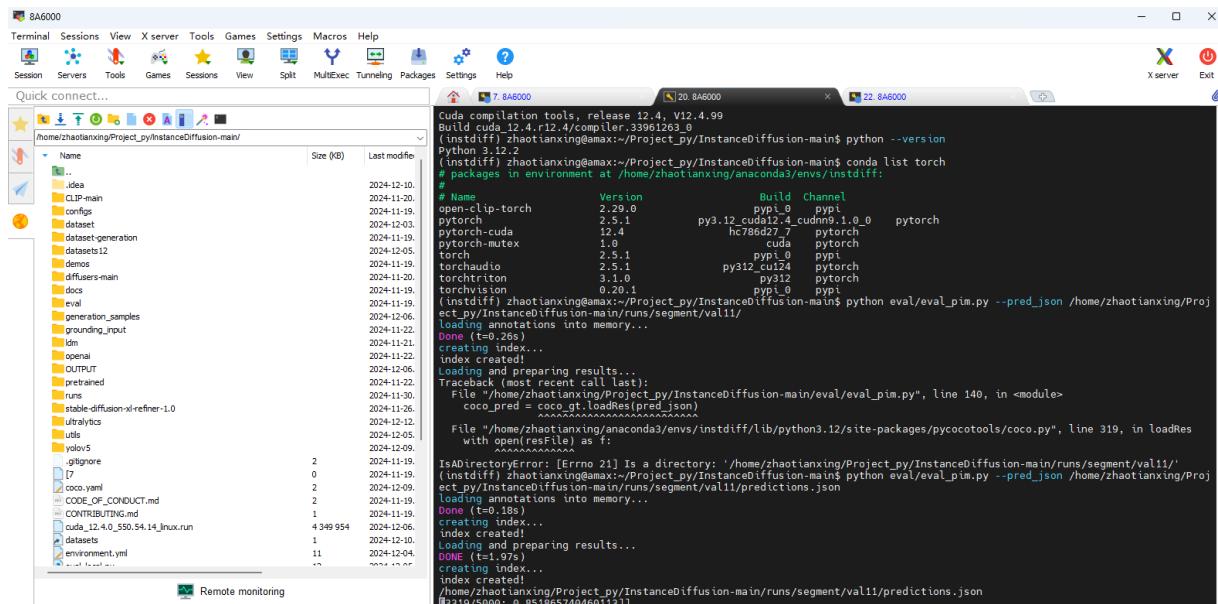


图 5. 操作界面示意

4.4 创新点

我在原论文所用数据集的基础上，尝试新增多模态的数据集，包含更多复杂的场景，样本中的实例也更加丰富。这样可以考察模型的泛化能力。由于目前研究的方向是 Story Visualization，因此我正在着手尝试将本文中创新性提出的三个重要模块进行优化和改进，以适配我正在进行的工作。因为我注意到了 ScaleU 在进行实例间的协调优化对 InstanceDiffusion 有一定贡献，使其能够精确控制每个实例的细节和全局一致性，而这正是 Story Visualization 任务中的需要考量的因素。

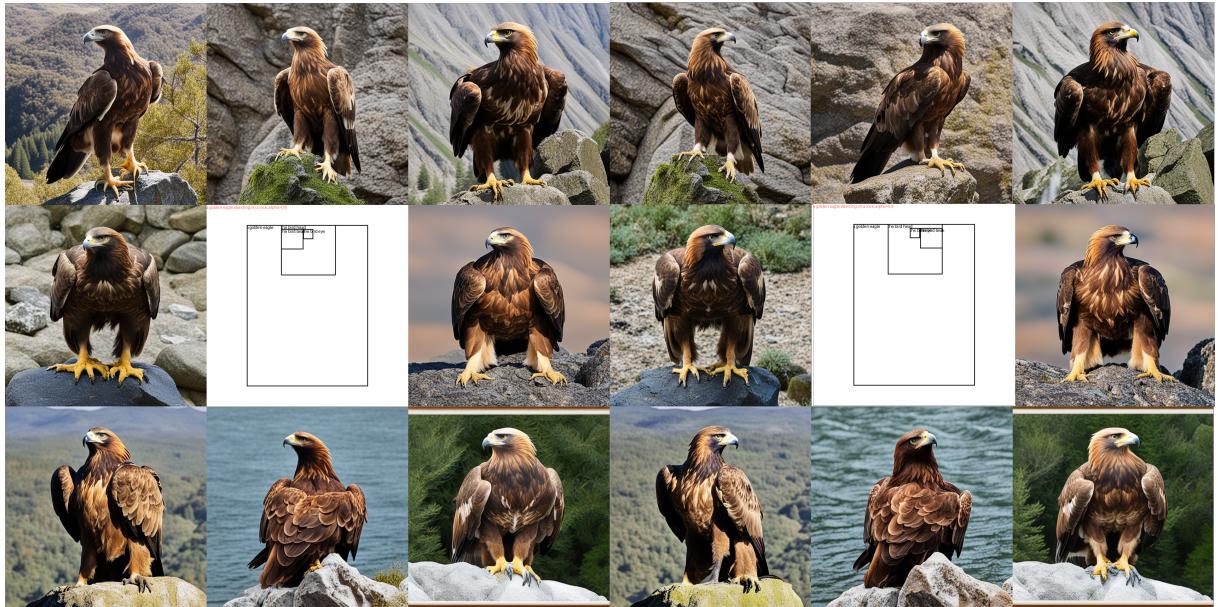


图 6. Eagle

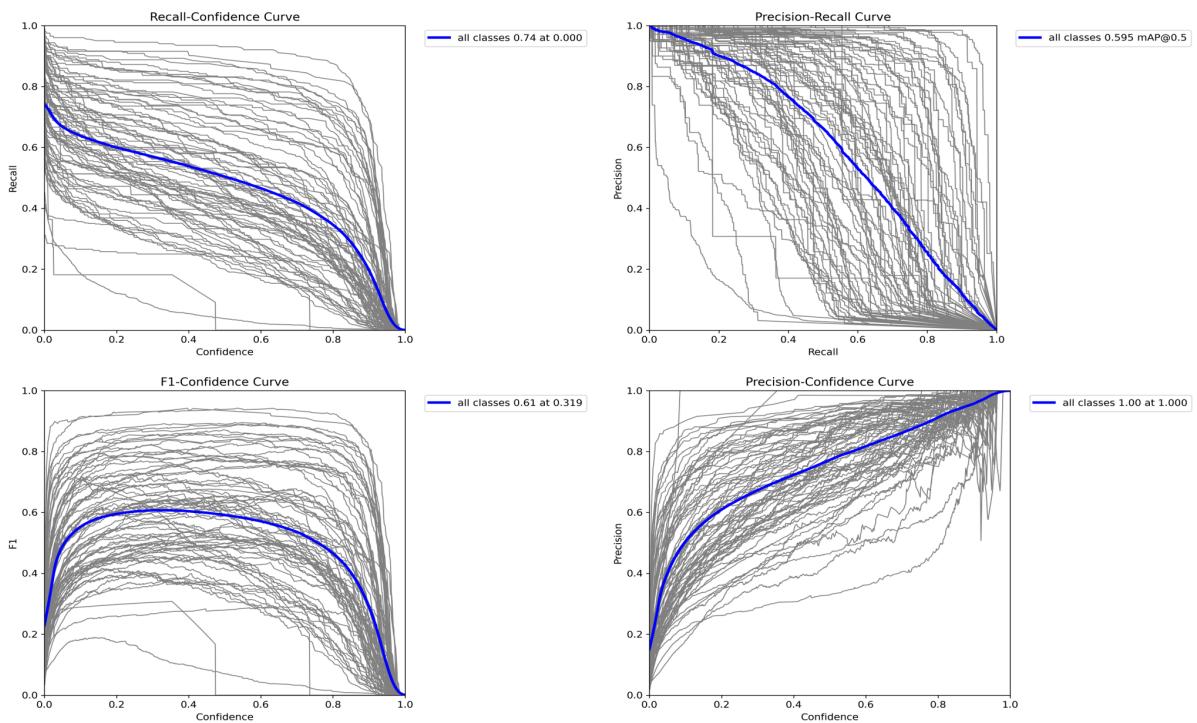


图 7. Curves

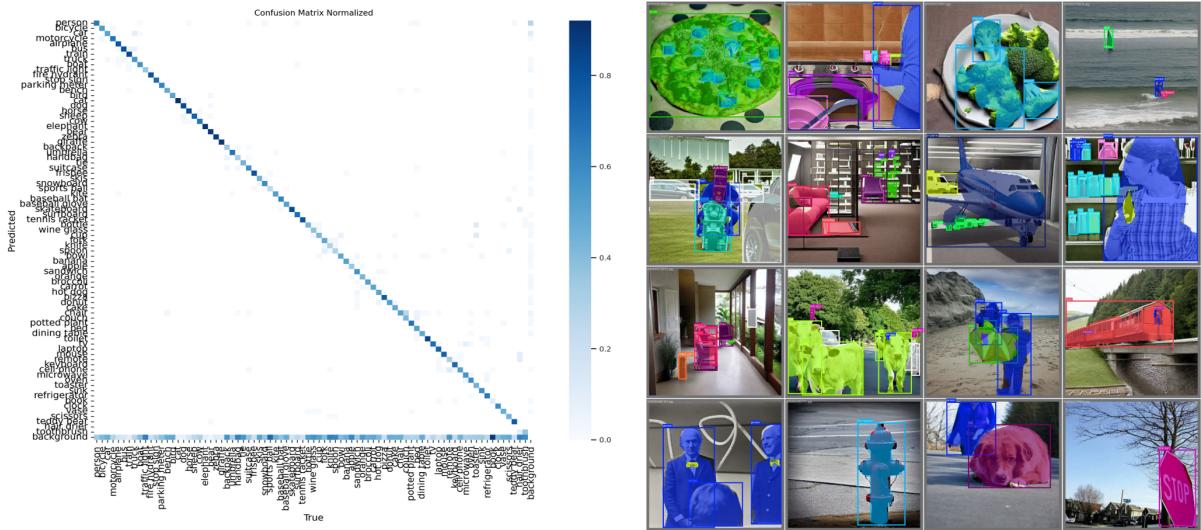


图 8. 混淆矩阵与可视化

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

如 6 所示，可以分为左半部分和右半部分，也就是两次生成，每次生成 8 张图片。右半部分 8 张图片是在左半部分的基础上仅改变实例的位置条件。中间白色区域提供了目标实例的位置信息以及其文本描述。文本描述是“a golden eagle standing on a rock”。位置信息用边界框进行表示，指定了鹰的头部、眼睛以及鹰喙的位置。

首先考察 InstanceDiffusion 在生成图像方面的多样性。每张图片的风格、背景和细节都有一定的差异。同一实例在不同的图片中呈现于不同场景中，包括高山、森林、湖等。同时实例本身的姿态、图片视角以及光影等方面具有一定差异。这些多样性得益于多实例采样器中每个实例的潜变量是独立生成的，体现了 InstanceDiffusion 在生成多样化图像方面的能力。

其次考察模型的实例级控制，在所有生成的图片中，目标实例的位置和大小均符合输入的边界框约束。从图 6 可以看出，右半部分和左半部分的区别仅在于实例的位置条件发生改变，对于场景等信息并没有发生大的变化。同时实例的特征表现也非常明显，比如图 6 中鹰的颜色是金色的。这说明该模型能够准确理解并执行文本描述和位置约束，体现了其在实例级控制上的优越性能。

最后考察其生成图像的质量。生成的实例质量较高，图 6 中鹰的姿态轮廓清晰，羽毛神态细腻。然而可以看到在第二次生成中，第二张图片鹰的头部朝向并未向右，因此这也是一个可以提升的地方。

如图 7 所示，(i)Recall-Confidence 曲线（左上）曲线显示了不同类别在不同置信度下的召回率变化情况。随着置信度的提高，召回率逐渐降低，这表明模型倾向于在高置信度下只预测更明确的结果，而可能漏掉一些边界不清晰的目标。蓝色曲线指示了平均召回率，在低置信度下较高，随着置信度增加而显著下降，说明模型在高置信度下仍存在部分漏检。(ii)Precision-Recall 曲线（右上）显示了各类别的精确率和召回率之间的权衡。大部分类别的曲线呈现递减趋势，表明提升召回率往往会以精确率的降低为代价。蓝色曲线为平均精确率，其 mAP@0.5 为 0.595，属于中等性能水平，说明模型在大多数类别上的精确率和召回率表现平衡，还有优

化空间。(iii)F1-Confidence 曲线（左下）显示当置信度为 0.319 时均值达到最大为 0.61，因此在这个置信度下，模型的平均精确率和召回率变现最佳。(iv)Precision-Confidence 曲线（右下）精确率随着置信度的提高而增加。随着置信度接近 1，模型的预测结果更加精准，但可能会降低召回率。蓝色曲线表示平均精确率在高置信度达 1，表明模型在最高置信度可以非常准确预测目标。

如图 8所示，左侧是归一化混淆矩阵，可以看出模型在部分类别上的分类性能较好，例如“cat”、“dog”、“horse”等，这表明这些类别的召回率较高，模型能够准确识别，也有不少类别之间存在明显的混淆。导致这一结果的原因可能是这些类别是数据样本分布不均。此外部分类别的召回率较低，可能是特征提取不足导致的。因此我认为可以通过增加容易混淆类别的数据样本加以采用适当的损失函数来进一步提高模型的性能。右侧的实例分割可视化展示了多种真实世界场景，包括室内和室外场景。图片中实例有多种类别，既有大物体也有小物体。可以看到部分实例的分割边界较为清晰，比如披萨上的西兰花、动物的轮廓灯。然而某些实例的分割的精度还欠缺，例如海边的人和物体等，但总体上模型表现还不错。同时，也可以从中注意到对于较小的实例分割掩码的边界不够完美，但边界框基本准确。其次对于多实例场景，可能出现错误分割或者目标重叠的情形。

6 总结与展望

InstanceDiffusion 被收录于 CVPR2024，这个工作可在文本到图像的生成过程中实现精确的实例级控制，在符合实例属性方面明显优于之前的所有工作，并支持各种位置格式-掩码、边界框、涂鸦和点。相对于传统的扩散模型，它针对复杂场景中的多实例生成问题，提出了模型设计提出了创新性的解决方案。通过引入 UniFusion 和 ScaleU 模块，使模型能够灵活地处理多种实例条件，并在生成过程中对不同实例进行分离建模。从而实现多实例的精准控制，确保生成图像语义全局一致，提高生成质量。并且通过在 COCO 数据集上的验证，并与 GLIGEN [24] 进行公平对比（模型训练设置以及训练集规模均与 GLIGEN 一致），展示了它在实例级控制上的优越性能。

其次，该论文的工作量也非常显著，包括了多个模块设计以及广泛的实验验证。将 InstanceDiffusion 和 GLIGEN、ControlNet 等先进方法进行全面对比，从生成质量、控制精度、适应性以及人类评估多方面进行对比，充分展示了其较先前工作的优势。还做了一系列消融实验证了新设计的 UniFusion 和 ScaleU 模块对模型最终性能的贡献。从实验的数据规模来看，不仅在 COCO 数据集上验证了模型，还扩展到了更多实例任务，表明了 InstanceDiffusion 的泛化能力。

InstanceDiffusion 通过引入新颖的设计提高了生成图像的质量，但也带来了更高的计算成本和训练难度。本次复现结果与论文中大体一致。同时在实验中也发现，相较于大物体，小物体的生成质量存在明显的下滑。在实例的纹理绑定方面，无论是 InstanceDiffusion 还是现有的其他方法，生成的效果仍然是不尽人意。对此，我认为未来改进方向：

- (1) 由于数据集中不同类别的样本占比不一，这也就导致样本量少的实例生成效果较差，因此我认为可以通过优化损失函数，即针对长尾数据分布，赋予稀有类别更高的损失权重。
- (2) 对 Diffusion 模型的优化，提高生成图像的细粒度，包括实例的纹理、边缘等特征，可以在模型中融入多尺度特征，引导生成更加精细的目标。通过多模态的条件控制，让模型能够

更精确地绑定纹理到对应的实例。

(3) 在扩散模型中集成一个纹理解码器，用于专门生成高精度的纹理贴图，然后再将纹理和实例进行绑定。但这也可能带来个问题，即纹理在物体上的过渡是否平滑，因此还需要对实例的边界进行约束。

这个工作对我的研究方向有一定启发，让我收获了在处理局部和全局一致性的有一种好的方法。后续将考虑融入 InstanceDiffusion 的条件生成方法，来确保生成帧间的可控性。

参考文献

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proc. Int. Conf. on Computer Vision*, 2023.
- [3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.
- [4] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.
- [5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.
- [7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.
- [8] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023.
- [9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [10] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffu-

- sion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.
- [11] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023.
 - [12] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.
 - [13] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022.
 - [14] Vudit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023.
 - [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
 - [16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
 - [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
 - [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
 - [19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
 - [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.
 - [21] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proc. Int. Conf. on Computer Vision*, 2023.

- [22] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, 2023.
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Euro. Conf. on Computer Vision*, 2014.
- [26] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.
- [27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [29] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention. *arXiv preprint arXiv:2212.00210*, 2022.
- [30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

- [34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [35] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023.
- [36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [39] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.
- [40] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [41] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023.
- [42] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023.
- [43] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1302–1310, 2020.
- [44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

- [45] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.
- [46] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, Weiming Zhang, and Nenghai Yu. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, 2023.
- [47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.