Inst-Inpaint：通过扩散模型指导物体移除

Ahmet Burak Yildirim、Vedat Baday、Erkut Erdem、Aykut Erdem 和 Aysegul Dundar

摘要—图像修复任务是指从图像中擦除不需要的像素，并以语义一致且逼真的方式对其进行填充。传统上，希望被擦除的像素通过二值掩码进行定义。从应用角度来看，用户需要为想要移除的物体生成掩码，这可能耗时且容易出错。在这项工作中，我们关注一种图像修复算法，该算法能够根据自然语言输入估计需要移除的物体，并同时将其移除。为此，首先，我们构建了一个名为**GQA-Inpaint** 的数据集，用于这项任务。其次，我们提出了一种新颖的图像修复框架，**Inst-Inpaint**，能够根据以文本提示形式给出的指令从图像中移除物体。我们设置了各种基于 GAN和扩散模型的基线，并在合成图像数据集和真实图像数据集上进行实验。我们使用不同的评估指标对方法进行比较，这些指标用于衡量模型的质量和准确性，并展示出显著的定量和定性改进。有关已发布代码和数据集，请参阅项目网页：http://instinpaint.abyildirim.com/

索引词——基于指令的图像修复，扩散模型。

arXiv:2304.03246v2 [cs.CV] 9 Aug 2023

1 引言

图像修复是指从图像中移除不需要的物体和/或填补缺失区域 [5], [26], [32]。这项任务极具挑战性，因为修复后的图像区域应与图像的其他部分协调一致。图像修复方法在填补缺失像素时，应考虑全局图像上下文。传统上，已有研究使用二值掩码标示缺失区域。与许多图像增强任务类似，近期的图像修复方法采用配备生成先验的基于深度学习的框架，并将图像修复任务建模为监督学习问题 [25], [46], [48]。也就是说，训练数据包括成对的输入图像和擦除后的图像，分别作为来源数据和目标数据；模型通常使用基于重建的目标函数进行训练。尽管近年来图像修复方法的性能持续提升，这些方法在推理阶段仍需要用户指定掩码。掩码可以通过绘制粗细不同的笔刷笔触或方块来提供，但绘制这些掩码往往既费时又容易出错。

近年来，基于文本的图像生成和编辑受到了广泛关注 [4], [36])。尤其是在大规模图文数据上训练后，DALL·E 2 [34],Stable Diffusion [35], Imagen [37] 和 Parti [47] 等模型展现出卓越的生成能力，同时还能有效捕捉图像的组合结构。此外，使用文本可以更自然、更便捷地控制图像。值得注意的是，研究人员还展示了，只需在模型中额外加入二值掩码作为条件，并指明要在掩码区域内填充什么内容，就能轻松将这些方法应用于图像修复。不过，这种方法更适合调整物体的属性或添加新的场景元素，却未必适用于移除物体。例如，请参阅 Blended Latent Diffusion 模型 [2]。

在这项工作中，我们提出了一项新任务，并将其称为 指令式图像修复：只通过文本指令指定需要移除的物体，无需二值掩码（见图 1）。为了解决这一问题，我们基于 GQA 数据集 [18], 首先利用数据生成流程创建一个新的基准数据集，并将其命名为 GQA-Inpaint。其次，我们设计了一个单阶段深度图像修复网络 Inst-Inpaint，可根据以文本提示形式给出的指令移除图像中的物体。与现有的图像修复方法相比，我们的模型无需掩码，也不会显式预测掩码。总之，我们的贡献如下：

- 我们提出了一种端到端的图像修复框架Inst-Inpaint，仅基于文本指令即可执行给定图像中物体的移除。我们方法的一些示例输出如图所示1。
- 我们生成了一个真实图像数据集GQA-Inpaint，用于训练和评估针对所提出的指令式图像修复任务的模型。在此，我们使用GQA 数据集 [18]该数据集最初是为视觉推理而提出的，并利用提供的场景图，通过采用最先进的实例分割和图像修复方法来生成配对训练数据。
- 我们进行了大量实验，以展示我们框架的有效性。为了比较结果，我们使用多项指标评估了各种基线方法，其中包括一种新颖的基于 CLIP [33] 的图像修复评分。我们

- *A. B. Yildirim* 和 *A. Dundar* 任职于土耳其安卡拉比尔肯特大学计算机科学系，邮箱为 *a.yildirim@bilkent.edu.tr* 和 *adundar@cs.bilkent.edu.tr*
- *V. Baday* 和 *E. Erdem* 任职于土耳其安卡拉哈杰特佩大学计算机科学系。
- *A. Erdem* 任职于土耳其伊斯坦布尔科奇大学计算机科学系。



图 1：**指令式图像修复**。我们提出了一种新的图像修复方法 **Inst-Inpaint**，该方法以图像和文本指令作为输入，能够自动移除文本中提及的物体，无需像传统图像修复方法那样由用户提供二值掩码来定位感兴趣区域。我们展示了该方法在一个新的真实图像修复基准数据集 **GQA-Instruct** 上训练和测试的示例结果。该数据集是我们为所提出的指令式图像修复任务构建的。

在基于文本的图像修复方法方面，取得了显著优于当前最先进技术的效果。

2 相关工作

基于扩散的生成模型。利用大规模图文配对数据集的扩散模型，在图像合成方面取得了令人瞩目的成果 [1], [3], [31], [34], [37]。扩散模型的图像合成能力已扩展至视频生成 [38], [43]和图像编辑 [1],[6], [15], [30], [39], [40] 任务。我们的工作针对基于文本指令的图像修复任务，与图像编辑任务的关系更为密切。SDEdit 是最早使用扩散模型编辑真实图像的工作之一 [30]。它对经过笔刷笔触修改的输入图像添加噪声，再对其进行去噪。由于经历了加噪和去噪步骤，无法严格保留图像内容。Prompt-to-prompt 编辑 [15]展示了通过对注意力机制进行操作，可以利用文本提示编辑生成的图像。Instruct- Pix2Pix [6] 利用 Prompt-to-prompt 生成一个数据集，并微调扩散模型，使其能够通过配对数据编辑图像。

图像修复。图像修复任务的常见设置包含一个用于标记已擦除像素的二值掩码 [25], [46]。原始图像像素会被移除，并根据这些掩码生成新的像素。该领域此前主要由生成对抗网络（GAN） [22], [25], [26], [32], [45], [46]主导。基于GAN的模型大多输出确定性结果，因为为提高稳定性，这些模型也会使用重建损失进行训练[25], [46], [48]。近期也有一些方法能够利用生成对抗网络实现多样化结果 [23], [51]。然而，这些方法是在诸如人脸之类的单一物体数据集上训练的。它们并未扩展到多样化场景的图像修复任务，因为生成对抗网络众所周知很难训练。

最近，基于扩散的模型也进入了这一领域，并取得了令人鼓舞的结果 [29], [36]。研究表明，a pre-训练好的无条件扩散模型可以通过 RePaint 模型 [29]对图像进行修复。RePaint 修改了去噪过- 程，使生成过程以未擦除的图像内容为条件。与我们工作同时进行的是 X-Decoder [53]，它可以同时处理输入图像和文本，以进行指代分割。X-Decoder可以根据提示对图像进行分割，并且与扩散模型结合后，可以擦除分割出的物体。

3 数据集生成

我们的目标是构建一个真实图像数据集，用于训练和评估基于指令的图像修补模型。目前存在一些可以调整用于此任务的数据集，例如 CoDraw [20] 和 CLEVR [19] 数据集。然而，这些数据集中的图像都描绘了由几何物体或剪贴画组成的合成场景。尽管这些相对简单的图像可视为构建模型的优秀工具，但在这些图像上训练的模型并不能直接应用于真实图像编辑。

为缓解现有数据集的这一缺点，我们在这项工作中提出构建一个新的图像修复数据集，其中包含复杂的真实图像，且无需使用任何人工标注。我们将这个数据集称为 *GQA-Inpaint*，因为我们选择基于 GQA 数据集 [18] 构建它。具体而言，GQA 数据集包含 8.5 万张真实世界图像及其对应的场景图。场景图通过以物体、属性和关系来表示图像，提供了图像的简化表示。场景图中的每个节点表示一个物体，其位置和大小由边界框指定。此外，每个物体还关联着若干属性，用于描述其颜色、形状或材质。连接节点（即物体）的边通常表示这些物体之间的空间关系。

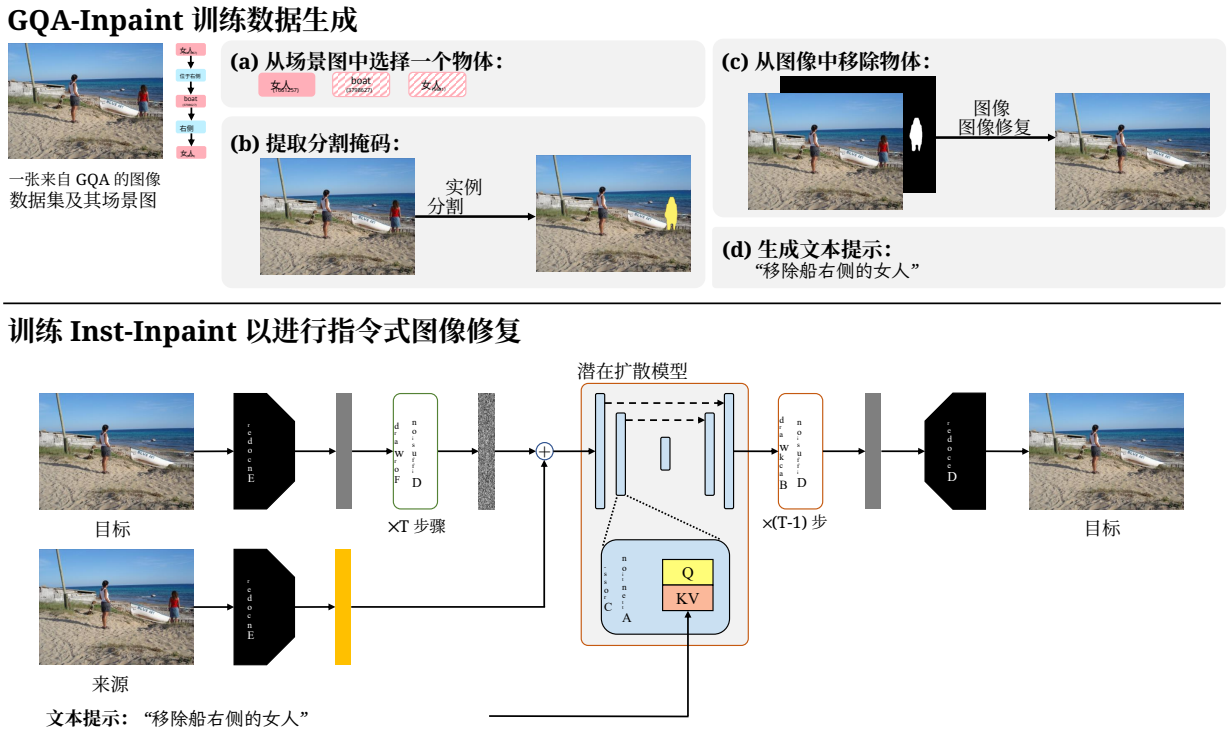


图 2： **所提出的 GQA-Inpaint 数据集和我们的 Inst-Inpaint 方法**。 我们的工作首先为所提出的 指令式图像修复 任务生成数据集。为创建输入/输出对，我们利用 GQA 数据集中的图像及其场景图 [18]。（a）我们首先选择一个感兴趣的物体。（b）我们执行实例分割，以定位图像中的物体。（c）我们应用最先进的图像修复方法擦除该物体。（d）最后，我们创建一个基于模板的文本提示，用于描述移除操作。因此，我们的 **GQA-Inpaint** 数据集共包含 **147165** 张独特图像和**41407** 条不同指令。在该数据集上训练的 **Inst-Inpaint** 模型是一种基于条件潜在扩散模型的基于文本的图像修复方法 [36] 无需用户指定任何二值掩码，即可在单步中执行物体移除，而无需像类似工作那样预测掩码。

我们提出的数据集的数据生成流程如图所示。2该流程高度依赖场景图表示。具体而言，图像修复任务中的物体选自场景图的节点。尽管 GQA 数据集中已提供物体的边界框，但我们发现，直接将其用作图像修复的掩码会导致结果不尽如人意。因此，为了获得更高质量的目标图像，我们提取每个物体的分割掩码，并用其从图像中移除物体。此外，我们还根据提供的场景图生成用作指令的文本提示。下面将介绍这些步骤。

从场景图中选择物体。在生成文本提示时，我们会从 GQA [18] 数据集的场景图中包含的物体中进行选择，主要依据以下两个问题：“从场景中移除该物体是否合理？”以及“移除操作是否会导致场景产生歧义？”。根据这些问题，我们可以将场景图中的物体分为两大类：(i)可以移除，且能够通过与其他物体的关系被唯一指代的物体，例如男人、船和风筝；(ii) 可以出现在指称表达中的物体，但移除它们并没有太大意义，例如墙和天空。我们手动为场景图添加属性，并使用“bidirectional”标记来定义这些类别。从从场景图的角度来看，可以说第一组中的物体具有双向关系，而第二组中的物体具有单向关系。例如，考虑图3所示的图像及其场景图。虽然当提及男人时，表达式中可能会出现墙左边的男人，但墙无法从场景中移除。另一方面，男人和汽车可以从场景中移除，因为它们的 bidirectionalflag 设置为 True。

部分场景图中的节点对应的不是单个物体，而是同一类别的多个物体，例如“苹果”、“自行车”、“窗户”等。我们会手动检测这些节点，并选择不将它们用作目标物体，而是将其视为可出现在指代表达式中的物体，用以指认感兴趣的物体。

我们还会根据物体的大小及其与其他物体之间的部分-整体关系，筛除某些物体实例。我们的标准如下：

- **物体尺寸**。我们不将面积大于整个图像面积一半的物体视为待移除物体。我们注意到，从图像中移除占据空间过大的大型物体，会导致不合理的图像修复结果。此外，我们也会舍弃过小的物体，即覆盖面积小于

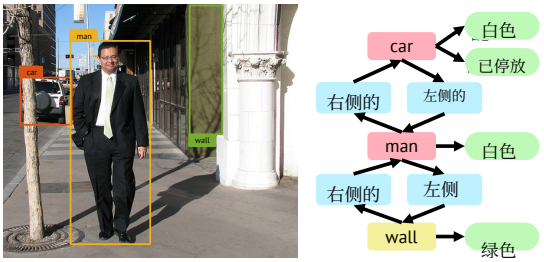


图 3：来自 GQA 数据集的示例图像 [18] 及对应的场景图。

整幅图像的 0.0025%。我们观察到，对如此小的物体进行图像修复后，生成的场景与原始场景非常相似。这些阈值由我们根据目视检查的结果通过经验确定。

- **部件与物体**。我们将那些可被视为其他物体的部件的物体类别排除在可移除物体列表之外。具体而言，这些物体可分为以下两类：

  – 物体的隐含部分。 例如，对于人体而言，腿、手臂和眼睛，对于汽车而言，车轮，以及对于猫而言，尾巴都属于此类别。对图像中的这些隐含部分进行图像修复，会导致场景不真实且视觉效果较差。– 物体穿戴的物品。例如，人穿着的夹克、裤子、衬衫、牛仔裤和鞋子 都属于此类别。与隐含部分类似，对这些可穿戴物品进行图像修复会产生不合理的结果，同时也会引发一些伦理方面的考量。

从场景图中选择关系。我们不使用场景图中存在的所有关系，而是移除出现频率低于 0.0001 的罕见关系。经过此筛选操作后，我们得到图中提供的关系。4此外，场景图中的一些物体没有与任何其他物体定义关系。对于这类物体，我们遵循一条简单规则，利用其空间位置生成关系。具体来说，我们沿 x 轴将图像划分为三个相等的部分，并按照自然的命名方式将这些部分命名为“左侧”、“中间”和“右侧”。我们以“位于 [位置]”的形式生成关系描述，其中 [位置] 指的是前述部分之一，该部分大体覆盖了感兴趣的物体。

提取分割掩码。我们使用 Detectron 获取物体的分割掩码 2 [44] 和 Detic [52] 框架。为覆盖广泛的物体类别，我们使用多个在 COCO [24]和 LVIS 数据集 [13]上训练的实例分割模型对物体进行分割。

- **Detectron2** [44]。我们使用Detectron2 ModelZoo 中提供的模型提取分割掩码，即Mask R-CNN [14], Cascade R-CNN [7], 和 Panoptic FPN [21]。这些模型训练所使用的类别标签通常与 GQA数据集中的类别标签不同。为了找到物体的正确类别标签，我们比较交并比

  预测边界框与场景图中所有物体边界框的（交并比）得分，并将该物体标记为具有最高交并比的类别标签

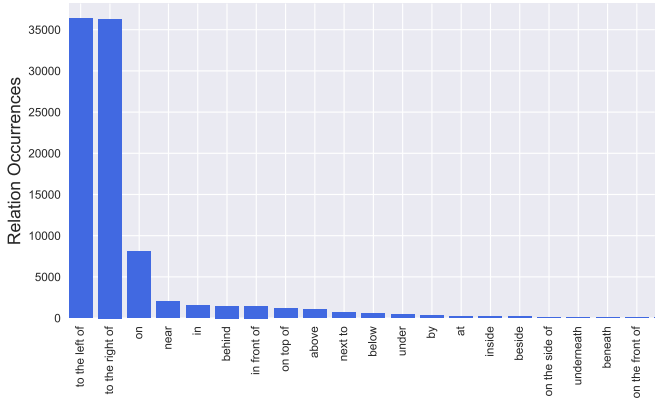


图4：所提出的 GQA-Inpaint 数据集中关系类型的分布（按出现次数排序）。

分数。

- **Detic** [52]。我们使用在 LVIS [13], COCO [24] 和 ImageNet-21K [9]数据集上训练的 SwinB-21K [27]模型来提取分割掩码。除了输入图像外，Detic 框架还需要一组可能的类别标签，以预测分割掩码。为此，我们采用 COCO 数据集中可用的物体类别列表，以及通过利用物体的属性和物体名称（例如“红色椅子”）以及仅使用物体名称（例如“椅子”）提取的自定义词汇表。为了获取词汇表与我们的词汇表不同的分割掩码，我们遵循上文介绍的 Detectron2 所采用的相同方法。

图 5 展示了上述方法和设置预测的实例分割掩码。通过结合不同模型的结果，我们为数量可观的物体类别获得了更准确的分割掩码。在这些由多种方法和设置生成的预测结果中，我们选择真实标注与检测边界框之间交并比较高的实例分割掩码，作为最终的物体掩码。

从图像中移除物体。 这一步的目标是从图像中移除物体。为此，我们采用最先进的图像修复方法 CRFill[49] ，因为它计算效率高且结果质量出色。在进行图像修复之前，我们首先使用 Cas- cadePSP [8] 优化预测掩码，然后使用结构元素为 11×11像素的形态学膨胀操作扩展边缘处的分割掩码，使掩码能够更好地覆盖物体。这些步骤如图 6所示。

生成文本提示。GQA 数据集包含物体之间的关系表示，如图中的场景图所示。2。如果所选物体是其类别中的单个实例，我们不使用任何关系，而是直接创建如下提示：“移除 *[物体]*”。如果图像中存在同一物体类别的多个实例，我们会根据场景图中的关系构建提示，以指定某个特定实例。对于要移除的物体，我们会使用其所有可用的关系。



图5：掩码提取方法比较。COCO数据集的词汇表与 GQA 数据集的词汇表不同。为了提取正确的分割掩码，我们结合使用 Detectron2的Model Zoo 中提供的预训练模型，并使用基于场景生成的词汇表，通过 theDetic2框架预测相应的类别标签。不同的设置为我们提供了更多选项，以选取最准确的实例分割掩码。



图 6：构建 GQA-Inpaint 时所考虑的物体移除操作中间步骤示意图。第一列展示原始图像，第二列展示由预训练实例分割模型检测出的分割掩码。第三列展示用于图像修复的物体掩码，该掩码首先经过 CascadePSP [8] 细化，然后通过膨胀操作进行扩展。第四列展示由 CR-Fill [49]得到的最终图像修复结果。

否则，我们使用图像相对位置。多个关系通过将它们全部与“男人左侧”和“桌子上”连接起来，组合成单一关系。例如，关系“男人左侧和桌子上”。除了关系之外，我们还使用待移除物体以及参与关系的物体的属性。与关系不同，我们不会组合所有可用的属性，而是随机选择一个属性作为数据增强方法。我们在图7中提供了数据集中的物体统计信息。

4 方法

为解决所提出的指令式图像修复任务，我们开发了一种名为 *Inst-Inpaint* 的新型条件扩散模型，该模型以图像和文本指令作为输入。训练该模型需要一个配对数据集，其中包含源图像和目标图像，以及相应的文本提示。我们基于潜在扩散模型 [36] 构建模型，以提高计算效率。具体而言，潜在扩散模型包括一个编码器E和解码器 D ，分别用于将图像投影到更低维的潜在空间，以及重建图像。它是一种变分自编码器，并作为第一步单独进行训练。

扩散过程 [17] 发生在潜在编码上，$z_0 = E(x)$，其中 x 表示输入图像。在每个时间步 t ，逐渐向 z_0 中添加噪声，因此经过T个步骤后， $z_T \sim N(0, I)$。扩散模型学习去噪自编码器经过训练，可以逆转这一过程，并预测其输入的去噪变体， z_t其中 z_t 是 z_0的一个含噪版本。这个作用于潜在编码的目标如下：

$$L_{LDM} := \mathbb{E}_{E(x),\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t)||] \quad (1)$$

其中， t 从 $1, ..., T$中采样，且 $\epsilon_\theta(., t)$ 是一个神经网络，用于根据时间步 t预测添加的噪声。与先前的工作 [36]，类似，我们采用 U-Net 架构。

如前所述，我们的指令式图像修复模型可以视为一个条件扩散模型。也就是说，我们使用两个额外的条件来训练模型：来源图像和文本提示。在训练过程中，目标图像经过编码器，得到 z_0 ，如图2所示。来源图像 s也使用第一阶段预训练的同一编码器进行编码。编码后的特征$E(s)$ 在每个时间步都与 z_t 拼接。对于文本条件，我们使用交叉注意力机制[41]。基于指令的物体移除目标如下：

$$L_{LDM} := \mathbb{E}_{E(x),\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t, E(s), T(i))||] \quad (2)$$

此处， i 指文本指令， T指变换器。编码器 E 在训练过程中保持固定，而 ϵ_θ 和 T 则一同优化。

实现与训练细节。我们训练了两个 Inst-Inpaint 模型，分别使用 GQA-Inpaint 和 CLEVR 数据集。我们在 CLEVR 数据集上训练模型的原因是为了将 Inst-Inpaint 模型与基于生成对抗网络的方法进行比较。

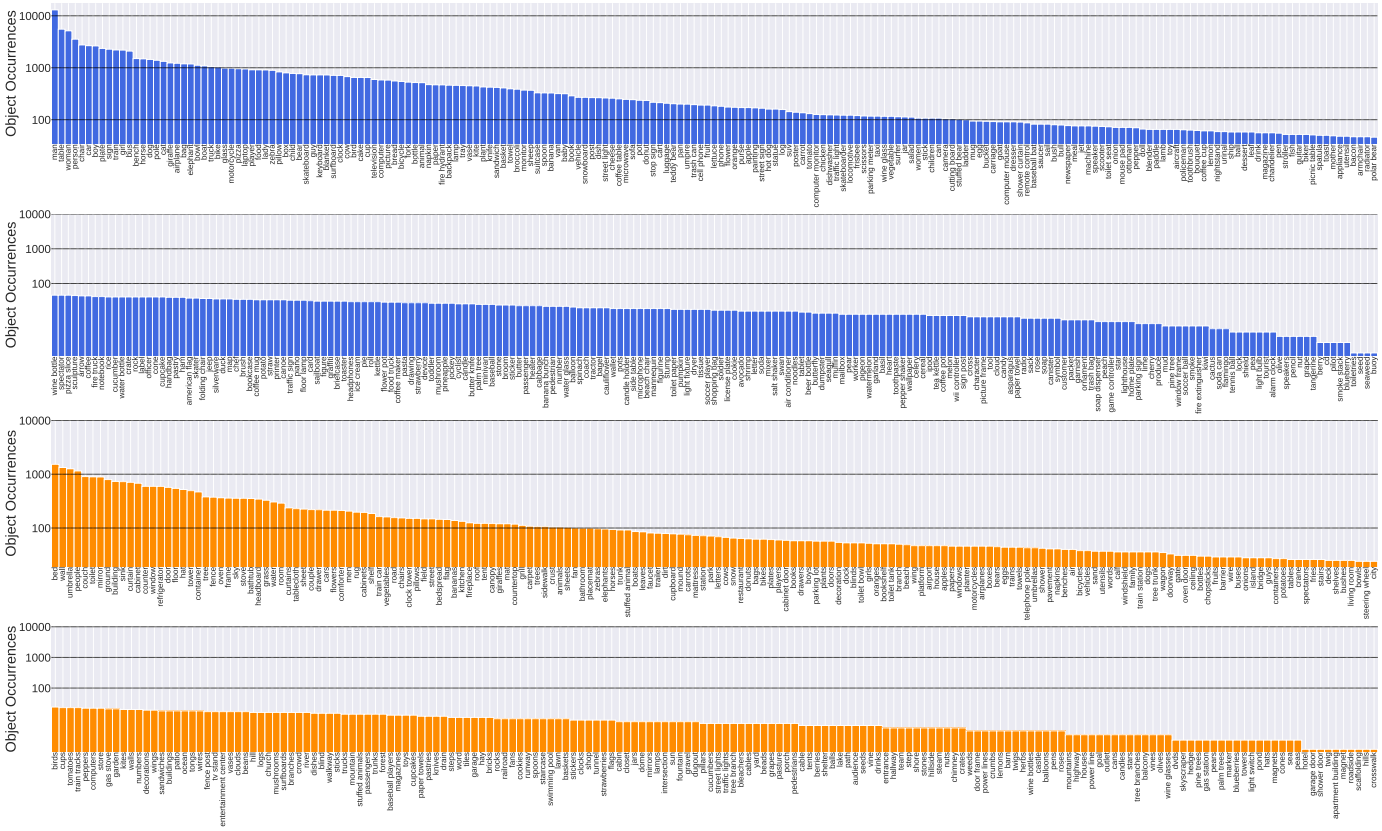


图 7：所提出的 GQA-Inpaint 数据集中各物体类别的分布，按出现次数排序并根据其作用分类。可被移除且可通过与其他物体的关系来指代的物体以蓝色显示，而可出现在指代表达中的物体以橙色显示。请放大查看文字。

由于我们未能在真实的 GQA-Inpaint 数据集上成功训练以生成对抗网络为基础的先前指令式图像擦除方法，因此我们在 CLEVR 数据集上与它们进行比较。关于这项比较的更多讨论见第 5.1 节。表 1 概述了超参数。GQA-Inpaint 模型使用预训练的 VQGAN 模型[1] 作为第一阶段编码器，而我们为 CLEVR 模型训练了一个经过 KL 正则化的自编码器。在两个训练阶段中，我们都对两个数据集先进行中心裁剪，再调整尺寸。

我们的模型在第二阶段采用两种条件信息：源图像和文本。对于源图像条件，我们将源图像的潜在表示与当前时间步的带噪图像拼接起来。潜在表示由第一阶段编码器提取，该编码器在训练阶段保持冻结。因此，U-Net 架构的输入数据所含的潜在通道数翻倍。对于文本条件，我们在文本嵌入与特定 U-Net 激活层之间应用交叉注意力。为此，与潜在扩散模型（LDM）的文生图模型一样，我们从头开始训练一个 Transformer 模型 [41]，并与 U-Net 一同输入 BERT [10] 词元，从图像修复指令中提取领域内语义表示。

5 实验

5.1 基线

我们将 Inst-Inpaint 与基于扩散和基于GAN的模型进行比较。作为基于扩散的对比模型，我们选用了 Instruct X-Decoder [53]，、CLIPSeg [28]，和 InstPix2Pix [6]。在这些模型中，Instruct X-Decoder 和 ClipSeg 会先分割出指令所指的物体。之后，它们可以与图像修复和编辑模型结合，实现不同的应用。我们对 Instruct X-Decoder 已发布的应用[2] 进行了实验。该应用结合了用于图像理解的 X-Decoder、用于语言理解的 GPT-3，以及用于图像生成的稳定扩散模型，并以零样本方式运行。需要指出的是，所有这些组件都在大规模数据集上训练，这些数据集远大于我们生成的 GQA-Inpaint 数据集。此外，我们还通过将 Instruct X-Decoder 的分割检测替换为 ClipSeg 的检测结果，对 ClipSeg 进行了实验。Inst-Pix2Pix 与我们的方法类似，都是端到端训练，根据指令转换图像。它是在作者生成的配对图像数据集上训练的。基于指令的配对图像由 prompt-to-prompt 方法生成 [15]。

目前，在大规模数据集上训练的图像生成和编辑方法通常采用扩散模型。我们还尝试将我们的模型与基于生成对抗网络的模型进行比较——

1. https://github.com/CompVis/taming-transformers

2. https://huggingface.co/spaces/xdecoder/Instruct-X-Decoder

	CLEVR	GQA-Inpaint
第一阶段编码器		
正则化类型	KL	VQ
z-形状	$16 \times 16 \times 4$	$32 \times 32 \times 4$
$\lvert\mathcal{Z}\rvert$	–	16384
训练轮数	400	预训练
自注意力残差	16	32
条件潜在扩散模型		
扩散步骤	1000	1000
噪声调度	线性的	线性
通道	64	128
深度	2	2
通道倍增系数	1,2,4	1、2、3、4
头数	4	8
批次大小	16	32
训练轮数	1000	300
学习率	2.e-6	2.e-6
文本条件交叉注意力		
分辨率	32, 64, 128	32, 64, 128
Transformer 深度	1	1
文本条件嵌入		
维度	512	512
Transformer 深度	8	16

表 1：在 CLEVR 和 GQA-Inpaint 数据集上训练的所提出 Inst-Inpaint 模型的超参数。



图 8：**基于 CLIP 的图像修复评分**。我们在评估中使用图像-文本相似度评分和零样本分类评分。评分说明请参阅第5.3 节。

，例如 TIM-GAN [50] 和 cMANIGAN [12]。TIM-GAN 通过对输入图像和文本进行编码，学习用于物体移除的掩码。编码后的特征以逐像素的方式与掩码相乘，并通过解码器重建图像。该网络使用对抗损失进行训练。此外，掩码预测还采用了重建损失。cMANIGAN 同样会预测掩码，并通过对抗损失和针对掩码的重建损失来学习编辑图像。
cMANIGAN 提出使用循环一致性损失：对图像进行两次编辑，第一次使用给定的指令，第二次使用其反向指令。这些模型在合成数据集上进行测试。我们在自己的数据集上训练这些模型，但未取得任何有意义的结果。为使研究更加完整，我们在合成数据集上将我们的方法与这些方法进行比较，即 CLEVR数据集。

5.2 数据集

GQA-Inpaint 数据集。 本数据集的构建过程已在第 3 节中详细说明。我们使用 44500 张独特的源图像和 88369 张独特的目标图像训练模型。由于我们会从源图像中选取不同的物体并生成指令，因此训练集包含 173715 个独特的源图像-目标图像-提示组合。我们使用 4811 张独特的源图像、9485 张独特的目标图像，以及 18883 个独特的源图像-目标图像-提示组合测试模型。
CLEVR数据集。 该数据集由 CLEVR 工具包生成[19]，[42]。数据集包含不同颜色和大小的三维合成简单形状。每个训练样本对都包含一张来源图像、一张目标图像，以及一段描述如何将来源图像修改为目标图像的文本。文本分为三类：添加、移除和更改一个物体。我们使用移除指令训练模型。为了扩大数据集规模，我们使用同时包含移除和添加指令的样本对。对于训练数据集，我们翻转包含添加指令的来源图像与目标图像，并相应地修改文本。该子集包含 5404 张训练图像和 2362 张测试图像。

5.3 评估指标

FID。 我们报告 Fréchet Inception 距离（FID）指标[16]，该指标通常用于通过比较目标图像分布与编辑后图像分布，评估生成图像的照片级真实感。对于我们的指令式图像修复任务，完全移除文本提示中指定的目标物体至关重要。FID 指标无法直接衡量移除操作的准确性，因此，为了对此进行评估，我们还使用基于 CLIP 的指标和 RelSim。

CLIP 距离。 CLIP 距离旨在评估指令中提及的目标物体是否被成功移除。我们使用与场景图节点关联的边界框来计算此指标。具体而言，我们提出的 GQA-Inpaint 数据集天然包含有关指令要求移除的物体边界框及其语义类别标签的信息。我们按照这些边界框的指示，从来源图像和修复后的图像中提取图像区域。随后，我们估算 CLIP 相似度 [33]，即这些提取区域与包含语义标签的文本提示之间的相似度，如图8所示。例如，我们使用文本提示“一张时钟的照片”来提取物体信息。在该示例中，这些相似度分数分别估算为 30。68 和 23。66。如果相似度值降低，则视为成功，如示例所示；我们报告其所占百分比。

CLIP 准确率。对于这项评估指标，我们使用 CLIP 作为零样本分类器。具体而言，我们根据目标物体的边界框，从来源图像中提取图像区域，并从数据集中已有的物体类别中找出该区域最可能对应的语义标签。接着，我们对修复后图像中提取的图像区域进行另一次预测。图中给出了 Top5 预测结果示例：8。我们期望执行指令所述的物体移除操作后，类别

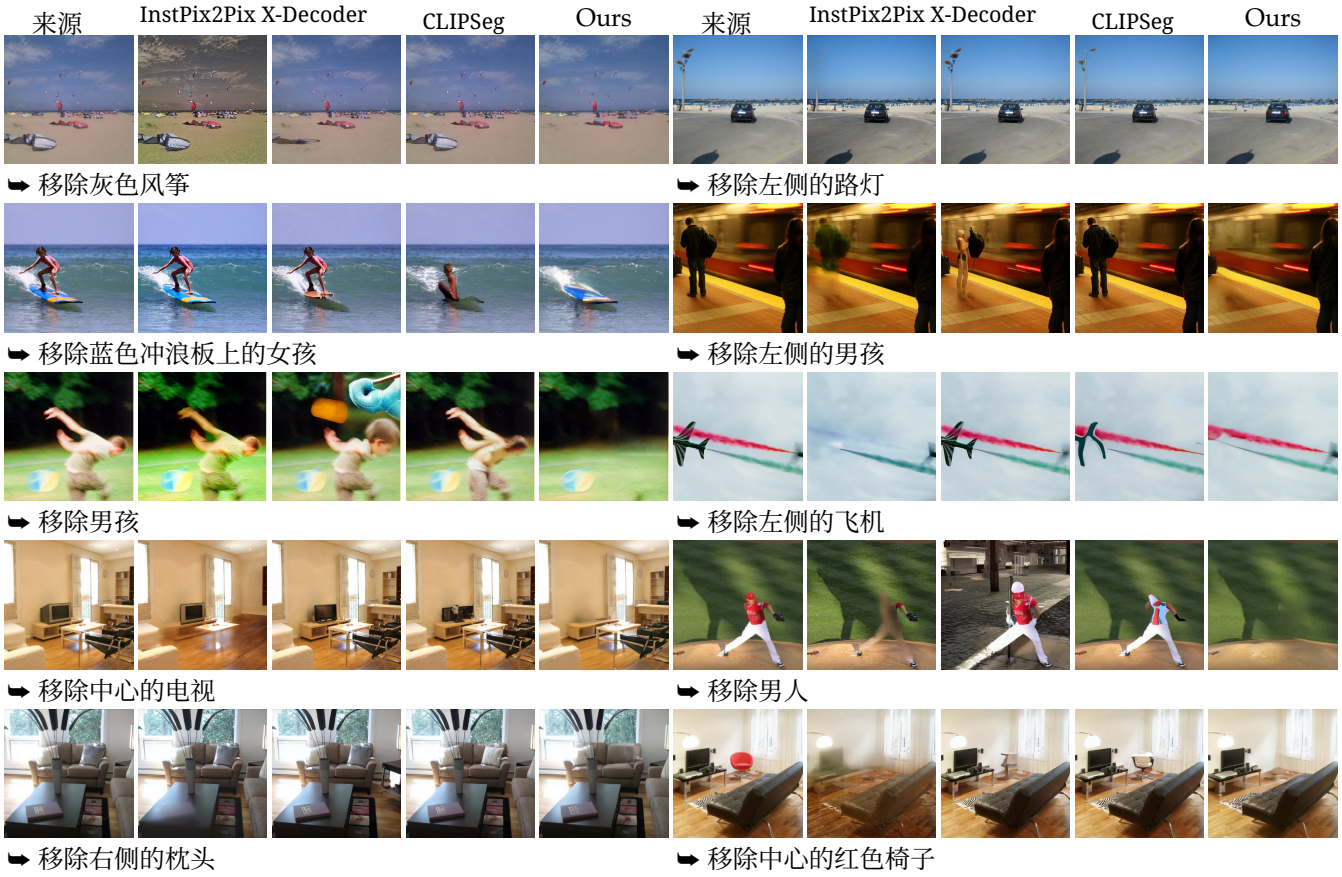


图9：我们的方法与竞争方法在 GQA-Inpaint 数据集上的定性结果。其他方法有时难以检测到正确的物体，也难以将物体完全擦除。另一方面，我们的方法取得了显著更好的结果。

方法	FID ↓	CLIP 距离 ↑	CLIP 准确率 ↑	CLIP ↑ 准确率（前5名）
X-Decoder [53]	6.360	72.2	62.6	41.5
InstPix2Pix [6]	9.972	56.8	33.5	11.8
CLIPSeg [28]	8.048	71.7	57.4	33.5
Inst-Inpaint（我们的方法）	**5.679**	**76.0**	**77.4**	**57.3**

表 2：我们的方法与竞争方法在 GQA-Inpaint 数据集上的定量结果。

预测结果会发生变化。我们主要关注 Top1 和 Top5 预测结果。如果基于来源图像预测出的类别未出现在修复后图像的 Top1 和 Top5 预测结果中，则视为成功。我们报告成功率。

RelSim。 此外，我们还使用了 GeNeVa 提出的关系相似度分数 [11]，该指标常用于 CLEVR数据集。它通过物体检测器和定位器，以定性方式比较物体的排列。我们使用 GeNeVa 发布并在 CLEVR数据集上微调的物体检测器。该检测器会构建修复图像和真实图像的场景图。该指标计算真实图像中的关系有多少出现在修复图像中。按照以往工作 [50]，我们在 CLEVR 上的评估中使用这一分数。不过，我们发现，这种评估方法难以用于真实图像，因为它需要高效的物体检测器。对于 GQA-Inpaint 上的实验，我们使用下文介绍的基于 CLIP 的度量方法。

5.4 实验分析

GQA-Inpaint 数据集上的结果。 我们在表格2和图9中分别展示了 GQA-Inpaint 数据集上的定量和定性结果。根据表格 2，我们的 FID 和 CLIP 得分均优于 X-Decoder 、InstPix2Pix 和 ClipSeg。如图9所示，InstPix2Pix 在许多示例中难以实现图像修复。在第一个示例中，模型没有移除灰色风筝，反而让图像变得更灰。有些示例中，它成功擦除了物体；而另一些示例中，它要么完全没有移除物体，要么只移除了部分物体。Instruct X-decoder 难以找到要移除的正确物体，在许多示例中也没有执行任何编辑，例如第一行第二个示例。在第一个示例中，它正确找到了要移除的物体，却在最终修复后的图像中生成了伪影。此外，它有时也确实会生成有趣的结果。例如，在许多示例中，尽管模型被要求根据背景填充修复区域，但移除电视、男孩和男人后，区域中却填入了同类别的物体。ClipSeg 的表现和失败情况也与 X-Decoder 类似。这种行

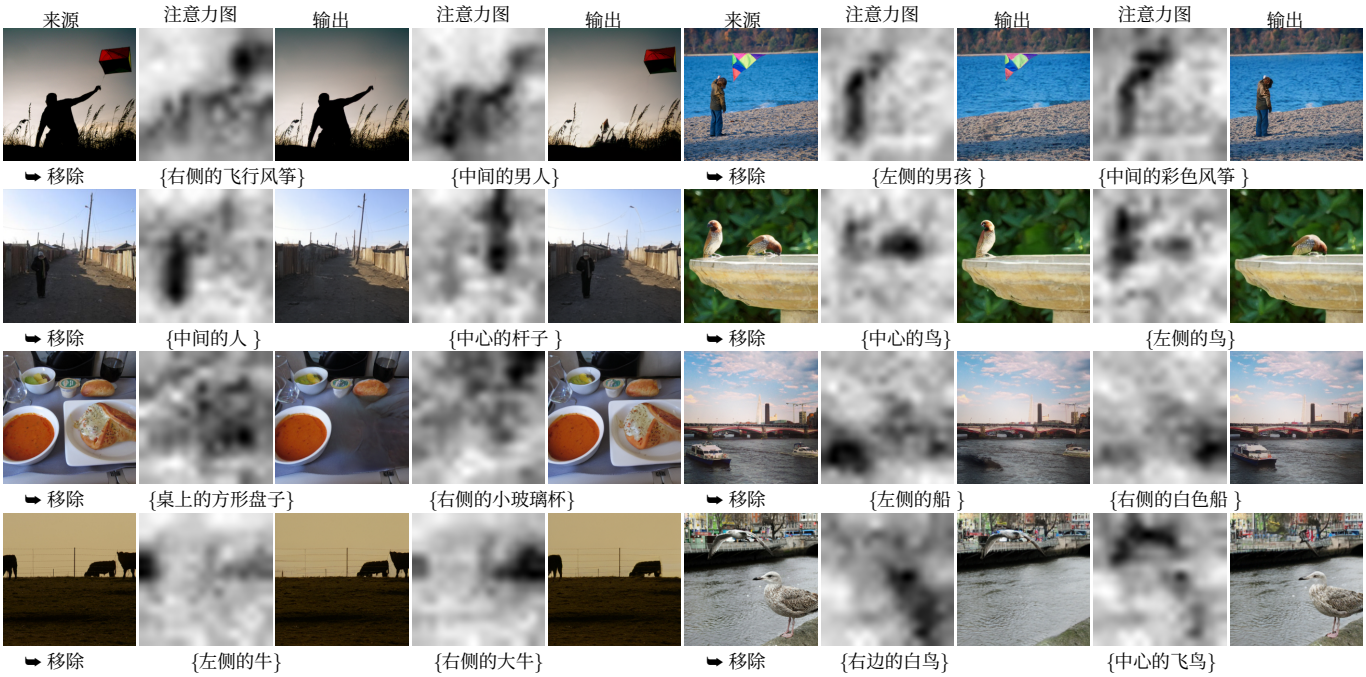


图 10：GQA-Inpaint 数据集上的其他基于指令的图像修补结果，以及我们方法的平均交叉注意力图。对于同一张输入图像，我们展示了基于两条不同指令生成的两种不同图像修复结果及其注意力图。

为导致 Instruct X-Decoder 和 ClipSeg 的 CLIP 得分较低。当物体未被编辑，或被替换为同类别物体时，CLIP 得分会受到负面影响。另一方面，如 X-Decoder 和 ClipSeg 的许多示例所示，如果图像没有进行任何编辑，则有利于 FID 得分。然而，我们的方法仍然取得了优于其他方法的 FID 得分，这是因为与竞争方法不同，我们的图像修复不会产生伪影。我们的 Inst-Inpaint 模型在具有挑战性的场景中成功擦除了物体。

分析。 我们在图中可视化了 GQA-Inpaint 图像样本的平均交叉注意力掩码。10 我们绘制了提示词所关注的像素。由于我们的任务是移除任务，我们发现注意力掩码聚焦于将被擦除的物体。如图 10所示，对于同一来源图像，我们的模型可以根据提示词擦除不同的物体，并且我们看到注意力图聚焦于我们想要移除的那些物体。

接下来，我们从对提示物体的分割效果来分析注意力图的准确性。我们的方法不会显式预测掩码，而是在端到端框架中对物体进行修补。不过，要擦除物体，网络需要隐式地检测待擦除的像素。我们在图10 中看到，注意力图能够检测出我们想要擦除的物体所在位置。我们希望将这些检测结果的准确性与其他基于提示预测掩码的方法进行比较，例如 CLIPSeg [28]和 X-Decoder [53]。为进行分析，我们训练了一个简单的 UNet，它以注意力图为输入并预测掩码。该 UNet 仅使用注意力图，不将提示或图像特征作为输入。这个 UNet 是单独训练的

CLIPSeg [28]	X-Decoder [53]	Inst-Inpaint（我们的方法）
40.6	48.7	53.1

表3：GQA-Inpaint 数据集上掩码预测的分割得分（交并比）。我们的方法不会显式检测语义图。我们训练了一个简单的网络，该网络基于我们的注意力图运行，以评估检测的准确性。

唯一目的是分析预训练的 Inst-Inpaint 的掩码预测。注意力图通过最近邻插值，从 16x16、8x8 和 4x4 的分辨率调整至 32×32 分辨率。我们只需将所有时间帧的注意力图拼接起来，从而得到一个包含 1280 个输入通道的张量。该网络包含 4 个下采样和 4 个上采样残差块层，通道数依次为 $\{32, 64, 96, 128, 128, 96, 64, 32\}$。最后，我们使用一个双线性上采样层，将分辨率提升至256×256。这是一个非常简单的设置，因为我们预计注意力图本身就能生成定义良好的预测。我们在表3中比较了掩码准确率。我们的方法取得了优于 ClipSeg 和 X-Decoder 的分数。

我们在图 11 中对 X-Decoder、ClipSeg 和我们的方法的掩码预测进行了可视化，这些预测来自以注意力图作为输入的UNet模型。正如根据图像修复结果所预期的那样，我们的模型更擅长预测正确的掩码。请注意，我们将 U- Net 设置为处理 32×32 分辨率，并且只在最后添加双线性上采样；使用功能更强大的解码器可以在掩码上获得更高的细节和更好的分数。为便于分析，我们将分割网络设置得较为简单。

与CLEVR数据集上的基于GAN的模型进行比较。

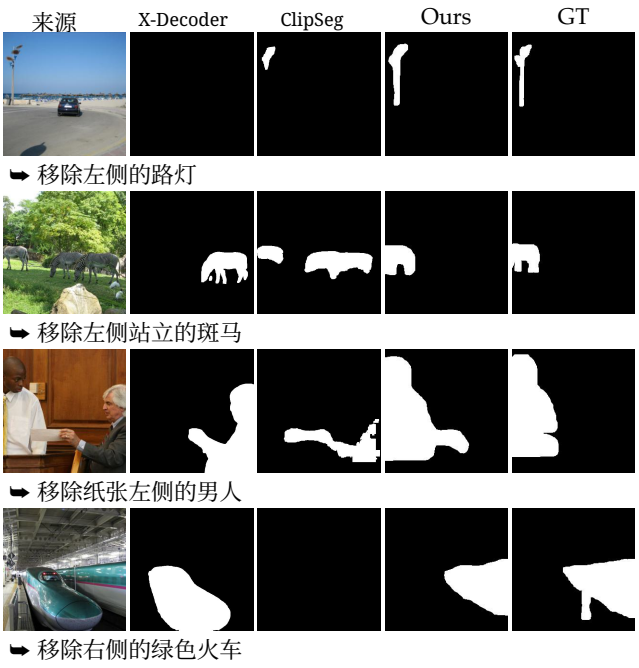


图11：我们在 GQA-Inpaint 数据集上的分割分析定性结果，以及竞争方法的掩码预测。

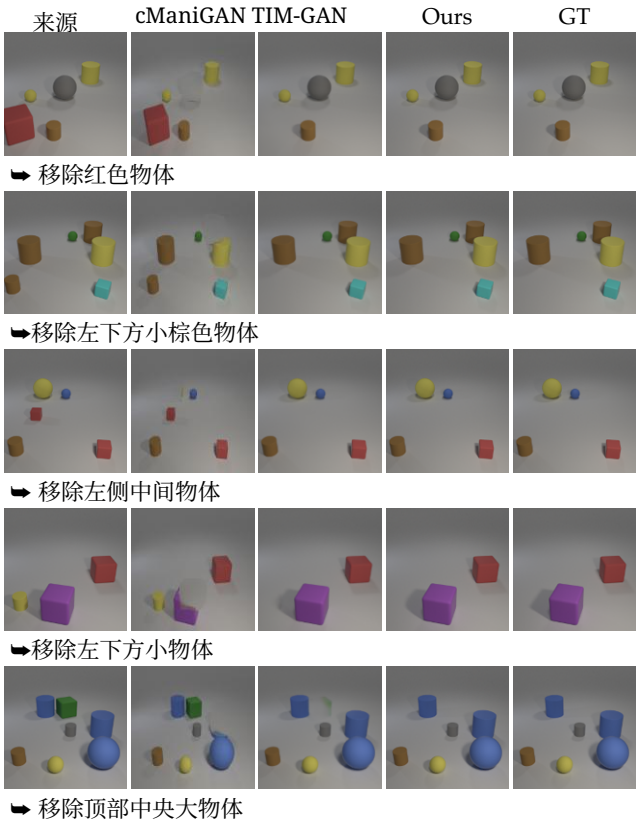


图12：我们的方法与竞争方法在合成CLEVR数据集上的定性结果。

方法	FID ↓	RelSim ↑
cManiGAN [12]	27.61	0.613
TIM-GAN [50]	0.558	0.919
Inst-Inpaint（我们的方法）	**0.453**	**0.914**

表4：我们的方法与竞争方法在合成CLEVR数据集上的基于指令的图像修复任务中的定量结果。

接下来，我们在更简单的CLEVR数据集上进行实验，并将我们的方法与基于生成对抗网络的方法在该数据集上的表现进行比较。我们在表 4 和图 12中分别给出了CLEVR数据集上的定量和定性结果。我们使用其发布的实现训练TIM-GAN。在作者的论文中，TIM-GAN在CLEVR数据集上进行了广泛测试，并且提供了预训练模型。然而，由于我们只关注移除指令，因此我们重新训练了他们的模型。对于cManiGAN，作者根据我们发送给他们的文本和输入图像，为我们提供了输出图像。

cManiGAN在CLEVR数据集上进行训练，但并非由我们训练。作者提到，他们使用的描述模板与我们的不同，可能会导致性能不佳。在我们的样本中，cManiGAN的视觉效果和定量结果确实较差。在许多示例中，待移除的物体未被正确检测出来，而且物体的图像修复也不够高保真。第一个示例中存在重影效果，并且在一些示例中物体仅被部分移除（最后4个示例）。我们的FID得分优于TIM-GAN，与其RelSim得分相当。从视觉上看，TIM-GAN获得的图像修复结果包含伪影，因为它并不总能完全移除物体。另一方面，我们基于扩散的Inst-Inpaint模型在这个简单的数据集上产生的伪影更少。请注意，与GQA-Inpaint数据集相比，这是一个相当容易的设置。我们分享这些结果，是为了与在该数据集上构建的基于GAN的模型进行比较，并展示Inst-Inpaint模型设置取得了比所提出的基于GAN的模型更好的结果。由于其他基于扩散的模型X-Decoder、CLIPSeg和InstPix2Pix是在大量真实图像上训练的，因此我们不将它们纳入此次比较。

5.5 模型局限性

我们以潜在扩散模型（LDM）作为基础架构来训练模型。该架构在潜在空间而非像素空间中执行扩散步骤。为此，模型使用自编码器来获取图像的潜在表示，并在第一阶段对其进行训练。对于 CLEVR 等简单数据集，重建图像与来源图像完全相同。然而，在 GQA-Inpaint 等复杂数据集中，具有复杂图案的场景（尤其是包含文字的场景）可能无法被自编码器良好地重建。即使我们的模型能够根据给定指令成功找到并移除目标物体，输出质量也可能因此较差。图形 1 - 移除彩色火车，可以看到，输入图像中的文字未能在输出图像中成功重建。我们可以构建更强大的自编码器或采用不同的优化方法来避免这些问题。我们希望发布数据集、基线模型和分析结果，能够推动这一领域的研究。

6 结论

在这项工作中，我们提出了一项新颖的基于指令的真实图像修复任务，其中仅通过文本指令指定不需要的物体，无需使用二值掩码。为此，我们开发了一个基于扩散模型的框架 Inst-Inpaint，仅凭文本提示即可从图像中移除物体。我们还创建了一个新的基准数据集 GQA-Inpaint，并为这项任务提出了新的评估指标。实验结果表明，我们的框架在这项任务上的表现优于现有方法。我们的工作展现了使用文本指令控制图像修复的潜力，这种方式更加友好且自然。我们将发布构建的数据集 GQA-Inpaint 和模型 Inst-Inpaint，以推动该领域的研究。

参考文献

[1] O. Avrahami、T. Hayes、O. Gafni、S. Gupta、Y. Taigman、D. Parikh、D. Lischinski、O. Fried 和 X. Yin。Spatext：用于可控图像生成的时空文本表示。arXiv 预印本 arXiv: *2211 14305*, 2022 2 。O. Avrahami、D. Lischinski 和 O. Fried。用于文本驱动自然图像编辑的混合扩散。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 18208 18218, 2022 1 – 页。Y. Balaji、S. Nah、X. Huang、A. Vahdat、J. Song、K. Kreis、M. Aittala、T. Aila、S. Laine、B. Catanzaro 等。ediffi：采用专家去噪器集成的文本生成图像扩散模型。arXiv 预印本 arXiv: *2211 01324*, 2022 2 。O. Bar-Tal、D. Ofri-Amar、R. Fridman、Y. Kasten 和 T. Dekel。Text2live：文本驱动的分层图像和视频编辑。在《计算机视觉–ECCV 2022：第17届欧洲会议，以色列特拉维夫，2022年10月23–27日，论文集》第XV部分，第707–723页。Springer， 2022 1 。C. Barnes、E. Shechtman、A. Finkelstein 和 D. B. Goldman。Patchmatch：一种用于结构化图像编辑的随机对应算法。《ACM图形学汇刊》，28 3 24, 2009 1 （）：。T. Brooks、A. Holynski 和 A. A. Efros。Instructpix2pix：学习遵循图像编辑指令。arXiv 预印本 arXiv: *2211 09800*, 2022 2, 6, 8 。Z. Cai 和 N. Vasconcelos。Cascade R-CNN：高质量目标检测和实例分割。《IEEE模式分析与机器智能汇刊》，43 5 1483 1498, 2019 4 （）：–。H. K. Cheng、J. Chung、Y.-W. Tai 和 C.-K. Tang。CascadePSP：通过全局和局部细化实现类别无关的超高分辨率分割。在 CVPR，2020 4, 5 。J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei。ImageNet：大规模分层图像数据库。在《2009 IEEE计算机视觉与模式识别会议》，第 248 255, 2009 4 – 页。J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova。BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本 arXiv: *1810 04805*, 2018 6 。A. El-Nouby、S. Sharma、H. Schulz、D. Hjelm、L. E. Asri、S. E. Kahou、Y. Bengio 和 G. W. Taylor。讲述、绘制并重复：基于持续语言指令生成和修改图像。在 IEEE/CVF国际计算机视觉会议论文集，第 10304 10312, 2019 8 – 页。W.-C. Fan、C.-F. Yang、C.-A. Yang 和 Y.-C. F. Wang。无需目标的文本引导图像操控。在 AAAI， 2023 7, 10 。A. Gupta、P. Dollar 和 R. Girshick。LVIS：大词汇量实例分割数据集。在 IEEE/CVF计算机视觉与模式识别会议论文集，第5356– 5364, 2019 4 页。K. He、G. Gkioxari、P. Doll´ar 和 R. Girshick。Mask R-CNN。在 IEEE国际计算机视觉会议论文集，第 2961 2969, 2017 4 – 页。A. Hertz、R. Mokady、J. Tenenbaum、K. Aberman、Y. Pritch 和 D. Cohen-Or。通过交叉注意力控制实现提示到提示的图像编辑。arXiv 预印本 arXiv: *2208 01626*, 2022 2, 6 。M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler 和 S. Hochreiter。遵循双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡。《神经信息处理系统进展》，30，2017。7

[17] J. Ho、A. Jain 和 P. Abbeel。去噪扩散概率模型。神经信息处理系统进展，33:6840– 6851, 2020 5 。D. A. Hudson 和 C. D. Manning。Gqa：用于真实世界视觉推理和组合式问答的新数据集。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 6700 6709, 2019 1, 2, 3, 4 – 页。J. Johnson、B. Hariharan、L. Van Der Maaten、L. Fei-Fei、C. Lawrence Zitnick 和 R. Girshick。Clevr：用于组合语言和基础视觉推理的诊断数据集。载于 IEEE计算机视觉与模式识别会议论文集，第 2901 2910, 2017 2, 7 – 页。J.-H. Kim、N. Kitaev、X. Chen、M. Rohrbach、Y. Tian、D. Batra 和 D. Parikh。CoDraw：作为基于接地的目标驱动通信测试平台的协作绘图。arXiv 预印本 arXiv: *1712 05558*, 2019 2 。A. Kirillov、R. Girshick、K. He 和 P. Doll´ar。全景特征金字塔网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 6399 6408, 2019 4 – 页。J. Li、N. Wang、L. Zhang、B. Du 和 D. Tao。用于图像修复的循环特征推理。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 7760– 7768, 2020 2 页。W. Li、Z. Lin、K. Zhou、L. Qi、Y. Wang 和 J. Jia。MAT：用于大孔洞图像修复的掩码感知变换器。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 10758 10768, 2022 2 – 页。T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar 和 C. L. Zitnick。Microsoft COCO：上下文中的常见物体。载于《计算机视觉——ECCV 2014：第 13 届欧洲计算机视觉会议》，瑞士苏黎世，2014 年 9 月 6—12 日，论文集，第 V 卷第 13 部分，第 740 755 2014 4 – 页。Springer，。G. Liu、A. Dundar、K. J. Shih、T.-C. Wang、F. A. Reda、K. Sapra、Z. Yu、X. Yang、A. Tao 和 B. Catanzaro。用于填充、图像修复和图像合成的部分卷积。IEEE模式分析与机器智能汇刊，2022 1, 2 。G. Liu、F. A. Reda、K. J. Shih、T.-C. Wang、A. Tao 和 B. Catanzaro。使用部分卷积修复不规则孔洞图像。载于欧洲计算机视觉会议（ECCV）论文集，第 85 100, 2018 1, 2 – 页。Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo。Swin Transformer：使用移位窗口的分层视觉变换器。载于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，2021 4 。T. Luddecke 和 A. Ecker。使用文本和图像提示进行图像分割。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 7086–7096 页，2022 年。6, 8, 9 A. Lugmayr、M. Danelljan、A. Romero、F. Yu、R. Timofte 和 L. Van Gool。RePaint：使用去噪扩散概率模型进行图像修复。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 11461–11471 页，2022 年。[30] C. Meng、Y. He、Y. Song、J. Song、J. Wu、J.-Y. Zhu 和 S. Ermon。SDEdit：使用随机微分方程进行引导式图像合成与编辑。载于国际学习表征会议， 2021 2 。A. Nichol、P. Dhariwal、A. Ramesh、P. Shyam、P. Mishkin、B. McGrew、I. Sutskever 和 M. Chen。GLIDE：利用文本引导的扩散模型实现照片级真实感图像生成与编辑。arXiv 预印本 arXiv: *2112 10741*, 2021 2 。D. Pathak、P. Krahenbuhl、J. Donahue、T. Darrell 和 A. A. Efros。上下文编码器：通过图像修复进行特征学习。载于 IEEE计算机视觉与模式识别会议论文集，第 2536 2544, 2016 1, 2 – 页。A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger 和 I. Sutskever。通过自然语言监督学习可迁移视觉模型。载于 M. Meila 和 T. Zhang（编），第 38 届国际机器学习会议论文集，机器学习研究论文集第 139 卷，第 8748–8763 页。PMLR， 18 24 2021 1, 7 – 7 月。A. Ramesh 、 P. Dhariwal 、 A. Nichol 、 C. Chu 和 M. Chen。使用 CLIP潜变量进行分层文本条件图像生成。arXiv 预印本 arXiv: *2204 06125*, 2022 1, 2 。R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer。使用潜在扩散模型进行高分辨率图像合成，2021 1 。R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer。使用潜在扩散模型进行高分辨率图像合成。

在 IEEE/CVF计算机视觉与模式识别会议论文集，第 10684 10695, 2022 1, 2, 3, 5 –页。C.Saharia、W.Chan、S.Saxena、L. Li、J. Whang、E. Denton、S. K. S. Ghasemipour、B. K. Ayan、S. S.Mahdavi、R.G.Lopes等。具有深层语言理解的逼真文本到图像扩散模型。arXiv 预印本arXiv: *2205 11487*, 2022 1, 2 。U. Singer、A. Polyak、T. Hayes、X. Yin、J. An、S.Zhang、Q.Hu、H. Yang、O. Ashual、O. Gafni 等。Make-a-video：无需文本-视频数据的文本到视频生成。arXiv 预印本 arXiv: *2209 14792*, 2022 2 。J.Singh、S.Gould 和L.Zheng。利用潜在扩散模型进行高保真引导图像合成。arXiv预印本 arXiv: *2211 17084*, 2022 2 。Y.Song、Z.Zhang、Z.Lin、S. Cohen、B. Price、J. Zhang、S.Y.Kim和 D. Aliaga。Objectstitch：生成式对象合成。arXiv 预印本 arXiv: *2212 00932*, 2022 2 。A. Vaswani、N.Shazeer、N. Parmar、J.Uszkoreit、L.Jones、A. N.Gomez、Ł. Kaiser 和 I. Polosukhin。注意力就是你所需要的一切。神经信息处理系统进展, 30, 2017 5, 6 。N. Vo、L.Jiang、C. Sun、K. Murphy、L.-J.Li、L.Fei-Fei 和J. Hays。组合文本与图像以进行图像检索——一次经验主义之旅。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 6439 6448, 2019 7 –页。J. Z.Wu、Y.Ge、X. Wang、W. Lei、Y.Gu、W.Hsu、Y. Shan、X. Qie 和 M.Z.Shou。Tune-a-video：用于文本到视频生成的图像扩散模型单次调优。arXiv 预印本 arXiv: *2212 11565*, 2022 2 。Y. Wu、A. Kirillov、F. Massa、W.-Y. Lo 和 R. Girshick。Detectron2。https://github.com/facebookresearch/detectron 2, 2019 4 。A. B. Yildirim、H. Pehlivan、B. B.Bilecen 和 A. Dundar。利用 GAN反演进行多样化图像修复与编辑。arXiv 预印本 arXiv: *2307 15033*, 2023 2 。J. Yu、Z. Lin、J. Yang、X. Shen、X.Lu 和 T. S. Huang。利用门控卷积进行自由形式图像修复。在 IEEE/CVF计算机视觉国际会议论文集，第 4471– 4480, 2019 1, 2 页。J. Yu、Y.Xu、J.Y. Koh、T. Luong、G.Baid、Z.Wang、V. Vasudevan、A. Ku、Y.Yang、B. K. Ayan、B. Hutchinson、W. Han、Z. Parekh、X. Li、H.Zhang、J. Baldridge 和 Y. Wu。扩展自回归模型以实现内容丰富的文本到图像生成。机器学习研究汇刊, 2022 1 。特色认证。Y. Yu、L. Zhang、H. Fan和 T. Luo。利用 GAN反演进行高保真图像修复。在《计算机视觉——ECCV2022：第 17 届欧洲会议，特拉维夫，以色列，2022 年 10月 23–27 日，论文集》，第 XVI 部分，第 242 258 2022 1, 2 – 页。Springer，。Y.Zeng、Z.Lin、H.Lu 和 V.M.Patel。Cr-fill：利用辅助上下文重建进行生成式图像修复。在 IEEE/CVF计算机视觉国际会议论文集，第 14164 14173, 2021 4, 5 –页。T. Zhang、H.-Y. Tseng、L. Jiang、W. Yang、H.Lee 和I. Essa。文本作为神经算子：通过文本指令进行图像操作。在第29 届 ACM国际多媒体会议论文集，第 1893 1902, 2021 7, 8, 10 – 页。S. Zhao、J.Cui、Y. Sheng、Y. Dong、X.Liang、E.I.Chang 和 Y.Xu。通过协同调制生成对抗网络进行大规模图像补全。arXiv 预印本arXiv: *2103 10428*, 2021 2 。X. Zhou、R. Girdhar、A.Joulin、P.Krahenbuhl和I. Misra。使用图像级监督检测两万个类别。在 ECCV, 2022 4 。X. Zou、Z.-Y.Dou、J.Yang、Z.Gan、L.Li、C. Li、X.Dai、H.Behl、J.Wang、L. Yuan等。像素、图像和语言的广义解码。arXiv预印本 arXiv:2212.11270，2022 年。2、6、8、9