arXiv:2302.11797v1 [cs.CV] 23 Feb 2023

面向零样本文本驱动图像编辑的区域感知扩散

NISHA HUANG，中国科学院大学人工智能学院；中国科学院自动化研究所模式识别国家重点实验室，中国FAN TANG，中国科学院计算技术研究所，中国WEIMING DONG，中国科学院自动化研究所；中国科学院大学人工智能学院，中国TONG-YEE LEE，台湾国立成功大学CHANGSHENG XU，中国科学院自动化研究所；中国科学院大学人工智能学院，中国



图 1. 所提出的区域感知扩散模型（RDM）的结果。文本遵循短语规则"A → B"，表示 RDM 将实体 A 转换为实体 B。

在文本描述的引导下进行图像操控近期受到了广泛关注。在本研究中，我们聚焦于在给定文本提示的引导下对图像进行区域编辑。不同于当前基于掩膜的图像编辑方法，我们提出了一种新颖的区域感知扩散模型（RDM），用于实体级图像编辑。该模型能够自动定位感兴趣区域，并根据给定的文本提示对其进行替换。为在图像保真度与推理速度之间取得平衡，我们通过结合潜空间扩散与增强方向引导，设计了密集扩散管线。此外，为了保留未编辑区域中的图像内容，我们引入了区域感知实体编辑，以修改感兴趣区域并保留

作者地址：Nisha Huang，huangnisha2021@ia.cn，人工智能学院、中国科学院大学；中国科学院自动化研究所模式识别国家重点实验室，中国；Fan Tang，tfan.108@gmail.com，中国科学院计算技术研究所，中国；Weiming Dong，weiming.dong@ia.cn，中国科学院自动化研究所；中国科学院大学人工智能学院，中国；Tong-Yee Lee，tonylee@mail.ncku.edu.tw，台湾国立成功大学；Changsheng Xu，csxu@nlpr.ia.ac.cn，中国科学院自动化研究所；中国科学院大学人工智能学院，中国。

允许免费制作本作品的全部或部分内容的数字或纸质副本，供个人或课堂使用，前提是副本不得用于或分发以谋取利润或商业利益，并且副本须附有本声明及首页上的完整引文。本作品中归他人而非 ACM 所有的组件，其版权必须得到尊重。注明出处后，允许摘录。若要以其他方式复制、重新发表、发布到服务器或分发至列表，则须事先获得特定许可和／或支付费用。请向 permissions@acm.org 申请许可。© 2023 美国计算机协会。提交至 ACM 的稿件

非关注区域。我们通过大量定性和定量实验，验证了所提出的 RDM 优于基线方法。结果表明，在视觉质量、整体协调性、非编辑区域内容保留以及文本-图像语义一致性方面，RDM 均优于以往方法。代码可在以下地址获取：https://github.com/haha- lisa/RDM-Region-Aware-Diffusion-Model。

CCS 概念： • **计算方法** → **计算机视觉任务**；图像操控。

附加关键词和短语：图像操控、文本引导、扩散

ACM 引用格式：
Nisha Huang、Fan Tang、Weiming Dong、Tong-Yee Lee 和 Changsheng Xu。2023。面向零样本文本驱动图像编辑的区域感知扩散。*ACM*多媒体计算、通信与应用汇刊1，1（2023年2月），17 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

在现实世界中，图像编辑的需求十分旺盛。然而，图像内容编辑软件不易上手，更适合专业人士使用。现有的图像编辑工作在可处理的输入图像和编辑操作方面受到限制。近年来，深度生成图像模型 [11, 14, 41]和视觉语言预训练模型 [38] 取得了巨大进展，使基于文本的图像生成与操控界面成为可能。

图像生成有几个主要分支，例如图像修复、图像翻译 [7], 风格迁移 [19],和图像操控 [1, 2, 36, 39, 47]。用于修改自然照片中物体风格和外观的计算方法已取得显著进展，使初学者也能实现各种编辑效果。尽管如此，值得注意的是，先前基于文本的图像操控方法要么不支持任意文本指令或图像操控 [5, 25, 34] ，要么只允许修改图像的外观属性或风格 [20, 30]。要控制修改区域，通常需要用户绘制区域进行指定 [1, 2, 5], ，这增加了操作的复杂性。本研究旨在消除上述所有限制，通过采用先进的图像生成技术，实现仅凭文本控制、开放式的图像内容修改。

近年来，多模态深度学习取得了长足进展，为机器实现跨模态通信与控制开辟了道路。应用广泛的大规模多模态预训练模型之一是对比语言-图像预训练（CLIP） [38] 模型，该模型使用 4 亿组文本-图像样本进行预训练。与此同时，各种新型图像合成方法 [15, 18, 25, 27, 37, 39] 展现了 CLIP 所涵盖的广阔视觉与语言领域的丰富内涵。然而，要操控任意真实图片中的现有物体，仍然颇具挑战。目前的主流技术将 CLIP 与预训练的 GAN生成器相结合，但输入图像的领域受到限制。近年来，人们对能够生成高质量、多样化图片的扩散模型产生了浓厚兴趣。另一方面，在像素级进行图像操控会导致生成时间延长，并消耗过多计算资源。

经过研究，我们发现，在预训练自编码器的潜在空间中应用扩散过程 [41]可以加快推理速度并减少计算资源的消耗。然而，先前的潜在扩散模型在生成图像真实感方面仍有不足。为了提升图像真实感并增强编辑结果与引导文本的一致性，我们在扩散的每一步都引入了无分类器引导 [17] 。基于这些发现，我们提出了一种用于零样本文本驱动图像编辑的高性能区域感知扩散模型（RDM）框架。具体而言，我们的方法能够编辑图像内容，使其符合任意给定的文本提示（如图 1 所示）。例如，给定一张狗的图像和定位文本：“一只狗”，我们的工作可

相应的编辑区域。然后，根据目标文本：“一只猫”，最终可以合成一幅高质量、逼真且多样化的图像。

为了让用户能够以简单直观的方式指定需要修改的区域和对象，以及需要创建的对象，我们提出了一个跨模态实体校准组件。该组件可以根据位置文本指导 t_1 定位并调整与文本相关的图像标记，从而正确识别需要修改的实体。我们观察到，将输入图像与掩码一同输入编码器后，在用于图像修改的扩散过程中，掩码之外的内容会出现意外变化。为了更大程度地保留无关内容，我们在每个扩散步骤中进一步执行扩散，将CLIP引导扩散结果与输入图像对应的带噪版本进行融合。此外，我们构建了相关损失函数，以保护非编辑区域并约束生成过程。我们将 CLIP梯度融入无分类器引导，使编辑结果更符合人类偏好，并使编辑区域生成的内容与目标文本指导 t_2的语义内容更加一致。总体而言，RDM 在潜在空间中实现了增强方向引导，从而生成快速、高质量、逼真且符合文本要求的编辑结果。

与以往方法的定量和定性比较表明，我们的方法能够通过文本更好地操控图像中的实体，同时不影响背景区域。如图 1、3 和4所示，在各种图像输入和文本描述的引导下，RDM 能够在对象内容变化方面生成逼真且高质量的结果。本工作的主要贡献总结如下：

- 我们提出了一种新颖的区域感知扩散模型（RDM），这是一种基于强大扩散模型的实体级零样本文本驱动图像编辑框架。
- 我们在扩散采样的每一步中引入空间位置掩码，并设计了非编辑区域保留损失函数，从而获得无拼接痕迹且未编辑区域得到良好保留的编辑结果。
- 我们在潜在空间中操控扩散步骤，并嵌入增强方向引导结构，以增强图像真实感，并提高控制文本与编辑结果之间的一致性。
- 定量和定性实验结果表明，在文本引导图像编辑的质量、真实性和多样性方面，RDM 优于基线方法，并取得了更出色的结果。

2 相关工作

本节回顾文本引导图像操控和扩散模型方面的现有研究，这些研究促使我们设计并实现了本应用。

2.1 文本引导图像操控

根据文本描述合成图像是一项雄心勃勃的任务，近年来取得了长足进展。最初的基于 RNN 的研究 [33]已被生成对抗方法超越。已有开创性研究将条件生成对抗网络应用于图像编辑 [12, 28, 35]。“Paint By Word” [5] 首次结合 CLIP [38] 与 StyleGAN2 [22] 和 BigGAN [6]，解决了零样本语义图像绘制问题。它只能改变图片的外观，例如颜色和纹理，却无法生成新的实体。ManiGAN [28]能够根据描述特定属性的文本，对图像的相应部分进行语义编辑，同时保留与文本无关的内容。然而，这类基于多模态生成对抗网络的方法限制了文本的表达能力。“Paint By Word” [5] 和 ManiGAN [28] 都受限于特定图像领域，不适用于开放域自然图像。

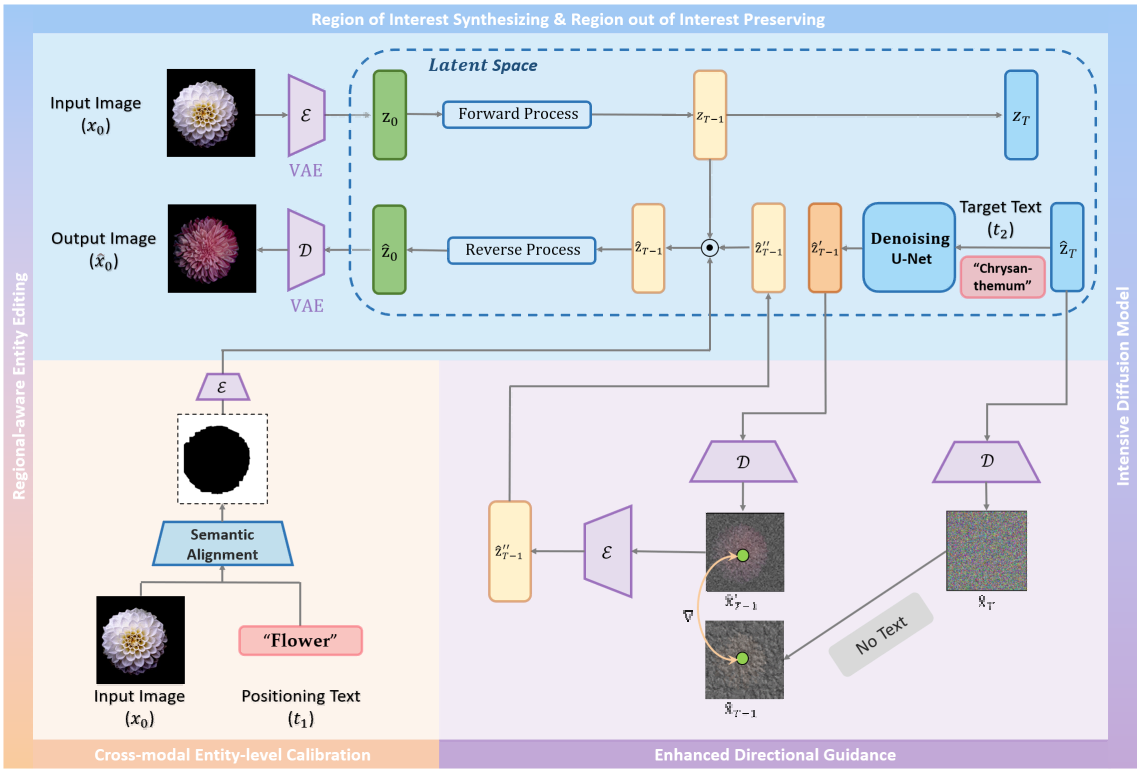


图2. 我们用于零样本文本驱动图像编辑的方法整体框架。

SDG [31] 和 DiffusionCLIP [25] 被提出用于利用扩散模型执行全局文本引导图像操作。GLIDE [36] 和 DALL·E 2 [39] 专注于文本驱动的开放域图像合成以及局部图像编辑。GLIDE 对其文本到图像合成模型进行微调，以实现图像修复。DALL·E 2 能够实现图像修复，但论文中对此缺乏讨论。两者都采用了将图像生成器和联合文本-图像编码器整合到架构中的思路。它们都包含使用大规模文本-图像对数据集训练的预训练模型，但都没有发布完整模型。后来，Blended Diffusion [2] 和 Latent Blended Diffusion [1] 被提出用于解决真实通用图像的局部文本引导编辑问题。然而，这些方法要求用户手动绘制额外的掩码，以指定图像的编辑区域，无法精确、自动地确定编辑区域。

2.2 扩散模型

扩散模型也称为基于分数的生成模型，是一种近年来发展迅速的强大生成模型。Sohl-Dickstein 等人提出了这一图像生成领域的新思路 [45]。当前研究 [11, 21, 46] 在高保真图像生成方面展现出令人惊叹的成果，甚至常常优于生成对抗网络。重要的是， [36, 39, 43]还具备出色的样本多样性，并能忠实覆盖学习到的数据分布中的各种模式。因此，扩散模型非常适合从复杂多样的数据中学习模型。

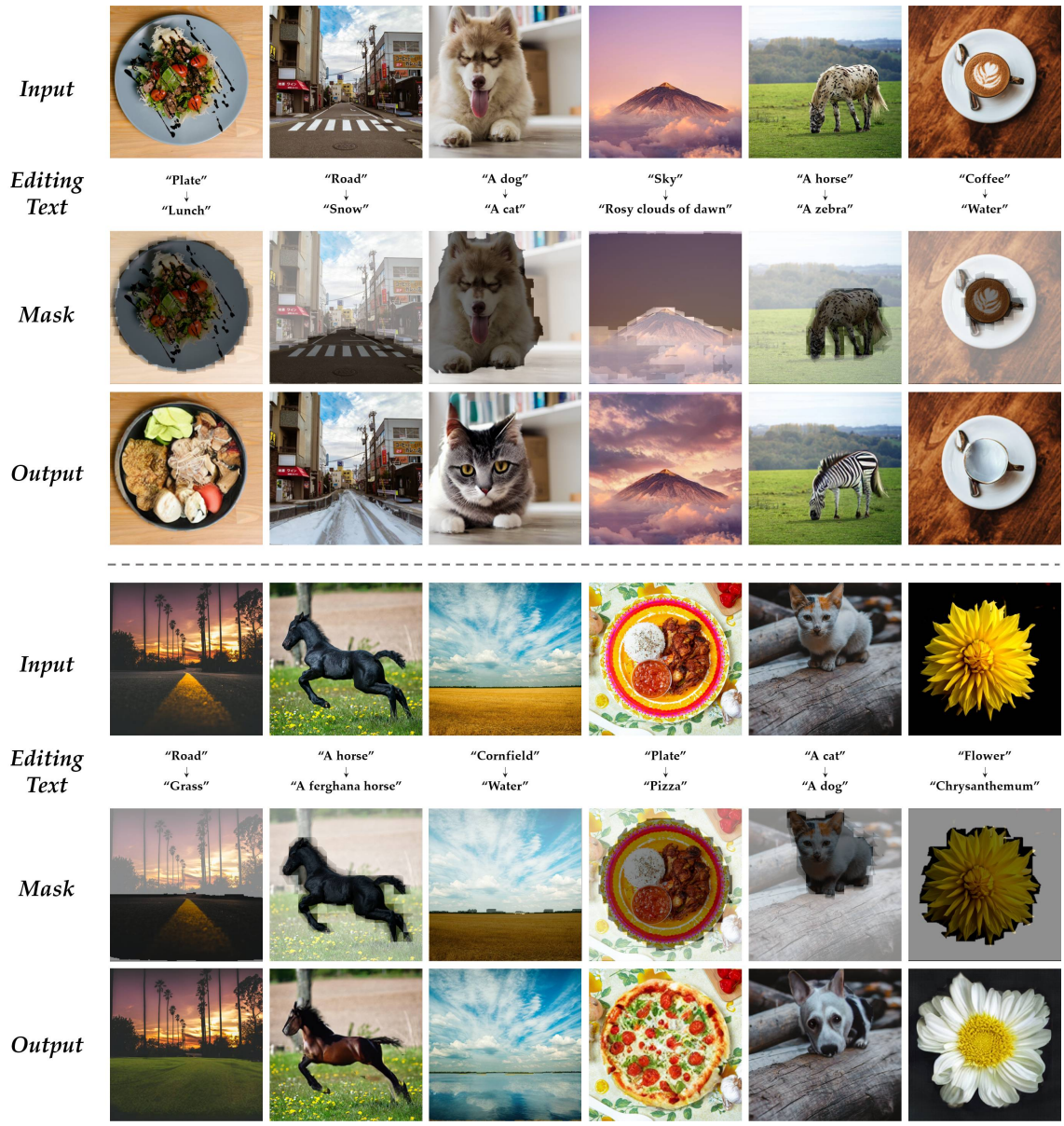


图 3. RDM 的更多操控结果。各行：不同的输入；各列：不同的编辑掩码和结果。

具体而言，扩散模型由一个前向过程和一个反向过程组成。前向扩散过程通过逐步扰动输入数据，将数据映射为噪声。反向过程则从纯随机噪声开始进行迭代去噪。扩散模型只需将随机采样的噪声输入经过学习的去噪过程，即可生成数据。扩散模型已应用于许多成功的领域，例如图像生成 [11, 21, 36, 39, 43, 46], 图像分割 [4], 图像到图像转换 [7], 超分辨率 [23, 41], 和提交至 ACM 的稿件

图像编辑 [1, 2, 36, 39]。Text2LIVE [3] 通过文本编辑现有物体的外观。它专注于生成一个叠加在原始输入图像上的编辑图层，而不是像我们这样移除或替换输入图像中的物体。

尽管上述方法在图像数据生成方面取得了前沿成果，但扩散模型的一个缺点是反向去噪过程缓慢。此外，传统扩散模型在像素空间中运行，因而会消耗大量内存。为加快采样过程并降低计算需求，研究者提出了潜在扩散模型（LDMs） [41] ，与基于像素的扩散模型相比，潜在扩散模型在这方面更具优势。LDMs 经过训练，可构建潜在视觉表示，并在低维潜在空间中执行扩散过程。 [41] 表明，该方法在各种计算机视觉任务上实现了新的最先进水平，并展现出极具竞争力的性能。然而，在图像保真度和文本-图像语义一致性方面，性能仍有提升空间。因此，我们设计了 RDM，在保留 LDM 速度优势的同时，兼顾图像质量和文本-图像对齐。

3 方法

3.1 概述

如图所示，本文提出的 RDM 是一个用于解决实体级零样本文本驱动图像编辑任务的框架。2我们的目标是通过一对文本提示（ x_0 ）来控制输入图像t_1， t_2，从而实现编辑。定位文本 t_1 用于确定待编辑实体的位置，而目标文本 t_2 用于生成新实体。在第3.2节中，我们将介绍密集扩散模型的具体组成。在第3.3节中，我们将说明如何通过文本实现区域感知实体编辑。

3.2 密集扩散模型

扩散模型 [45] 是一种可用于生成图像的生成器。扩散过程分为前向过程和反向过程：前者向输入图像 x_0添加随机噪声，后者则移除噪声并生成图像 $\hat{x}_0$。与传统扩散模型 [11, 16], 不同，我们不在像素级执行扩散过程。近期的一些研究 [1, 41] 表明，在潜在空间中执行扩散过程可以降低计算开销并加快采样过程。去噪 U-Net 学习移除噪声，并在经过 T 步去噪后生成输出图像 $\hat{x}_0$。然而，在潜在空间中执行扩散过程会损害图像生成质量和文本-图像一致性。为改善这一问题，我们进一步引入了增强方向引导组件。

潜在表示。 如上所述，我们在潜在空间中进行扩散步骤 [41]，以降低复杂度并实现高效的图像处理。我们使用变分自编码器 VAE [26]来实现感知图像压缩。扩散模型直接在低维潜在空间中运行，充分利用图像特定的归纳偏置。这样，底层自编码器主要由二维卷积层构成，并通过使用重加权界，进一步将目标聚焦于感知上最重要的信息位，这些信息位表示为：

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t\right)\|_2^2\right]. \tag{1}$$

我们模型的确定过程 $\epsilon_\theta(z_t,\ t)$ 由时间条件 UNet [42]实现。由于前向过程是固定的，训练期间可以方便地从 $\mathcal{E}$ 获取 z_t，而 $\mathcal{D}$ 则可以将样本从 $p(z)$ 解码到像素空间。

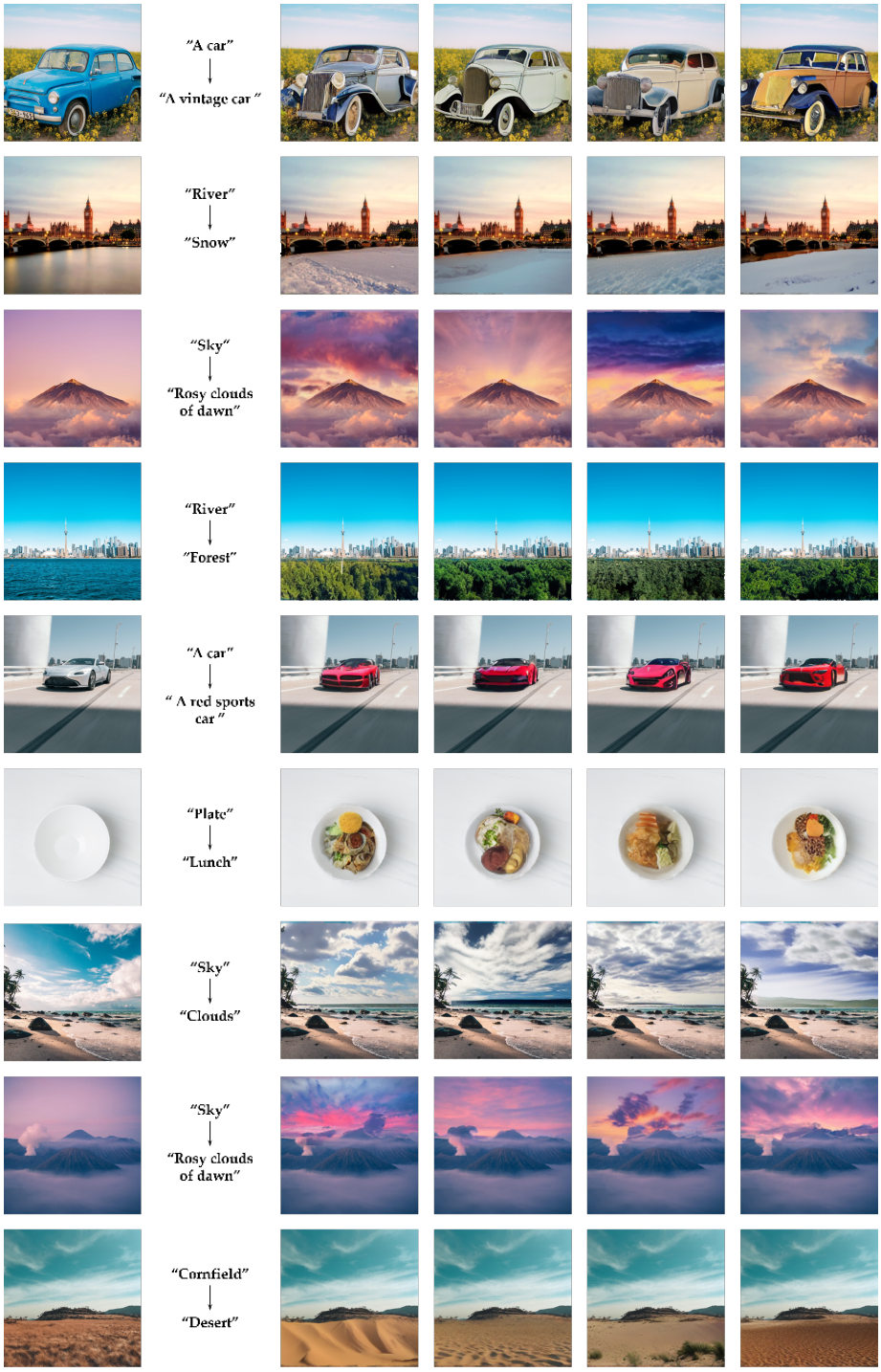


图4. RDM 的更多操控结果。

增强方向引导。为强化源区域遵循目标文本的编辑方向，我们尝试修改无分类器引导 [17] 以增强跨模态引导。这是一种无需训练单独分类器模型即可引导扩散模型的策略。通常，无分类器引导具有两项优势。首先，它无需依赖单独的（且可能较小的）分类模型的知识，而是允许单个模型在引导过程中利用自身经验。其次，当条件信息难以通过分类器进行预测时，它能简化引导过程。

为了提供无分类器引导，在训练过程中，类别条件扩散模型 $\epsilon_\theta(x_t \mid y)$ 中的标签 y 会被空标签 $\emptyset$ 替换。在采样期间，模型的输出会进一步沿 $\epsilon_\theta(x_t \mid y)$的方向延伸，并远离 $\epsilon_\theta(x_t \mid \emptyset)$:

$$\hat{\epsilon}_\theta\left(x_t \mid y\right)=\epsilon_\theta\left(x_t \mid \emptyset\right)+s \cdot\left(\epsilon_\theta\left(x_t \mid y\right)-\epsilon_\theta\left(x_t \mid \emptyset\right)\right). \tag{2}$$

推荐的引导尺度为 $s=5$。该方程受到分类器的启发，

$$p^i\left(y \mid x_t\right) \propto \frac{p\left(x_t \mid y\right)}{p\left(x_t\right)}. \tag{3}$$

真实分数的作用是表示梯度 ϵ^*，

$$\nabla_{x_t} \log p^i\left(x_t \mid y\right) \propto \nabla_{x_t} \log p\left(x_t \mid y\right)-\nabla_{x_t} \log p\left(x_t\right), \propto \epsilon^*\left(x_t \mid y\right)-\epsilon^*\left(x_t\right). \tag{4}$$

修改后的预测 $\hat{}\epsilon$ 随后用于引导我们生成目标文本提示 t_2，如算法 1所示:

$$\hat{\epsilon}_\theta\left(x_t \mid t_2\right)=\epsilon_\theta\left(x_t \mid \emptyset\right)+s \cdot\left(\epsilon_\theta\left(x_t \mid t_2\right)-\epsilon_\theta\left(x_t \mid \emptyset\right)\right). \tag{5}$$

3.3 区域感知实体编辑

跨模态实体级校准。 为了基于定位文本 t_1生成二值分割掩码 m，我们设计了一个跨模态实体级校准模块，由一个预训练的CLIP模型和一个精简的条件分割层（解码器）组成。首先，将定位文本 t_1 输入CLIP文本Transformer，以获取条件向量。受逐特征变换 [13, 32], 的启发，条件向量用于调制解码器的输入激活。这使解码器能够将CLIP中的激活与分割输出关联起来，并告知解码器分割的目标。将输入图像 x_0 输入CLIP视觉Transformer，以获取 $\mathbb{R}^{W \times H \times 3}$。随后，在每个Transformer块之前，将从层 $S=[3,7,9]$ 提取的激活与解码器在嵌入维度 $F=64$ 上的内部激活相加。此外，CLIP ViT-B/16使用的令牌块大小为：$P=16$。解码器通过对其Transformer（最后一层）的令牌进行线性投影，生成二值分割:

$$\mathbb{R}^{\left(1+\frac{W}{P} \times \frac{H}{P}\right) \times F} \mapsto \mathbb{R}^{W \times H}. \tag{6}$$

为了将 CLIP 的能力与分割结果关联起来，我们采用通用二值预测设置。对于空间掩码 m，我们对二值分割进行阈值处理，阈值 K 的范围为 0 到 255，通常取 150。

感兴趣区域合成。为了使感兴趣区域能够根据文本提示进行编辑，我们利用预训练的 ViT-L/14 CLIP [38] 模型进行文本驱动的图像内容操控。扩散过程中，去噪图像$\hat{}x_t$ 的 CLIP嵌入与文本提示t_2 的 CLIP嵌入之间的余弦距离可用于定义基于CLIP的损失，或$\mathcal{L}_{CLIP}$。目标文本提示 t_2 被嵌入到嵌入空间中，记为 E_L。用于噪声图像的时变图像编码器记为 E_I。我们定义提交至 ACM 的稿件

基于余弦距离的语言引导函数，用于衡量嵌入 E_I 与 E_L 之间的相似程度。文本引导函数可定义为：

$$\mathcal{L}_{CLIP}\left(\hat{x}_t, t_2, m\right) = E_I\left(\hat{x}_t \odot m\right) \cdot E_L(t_2). \tag{7}$$

上述过程不受任何额外的非编辑区域限制。尽管 L_{CLIP} 是在正在编辑的区域内进行评估的，但它也会影响非编辑区域。为了解决这一问题，我们在下文给出等效的方法。

感兴趣区域外内容保留。上述过程中没有采用非编辑区域保留（NERP），该过程从各向同性高斯噪声开始。因此，即使 $\mathcal{L}_{CLIP}$ 是在掩码区域内进行评估的，它仍会影响整幅图像。为缓解这一问题，我们将掩码 m 编码到潜在空间中，得到 $m_{\text{潜}}$ 变量，并按如下方式将其融入扩散过程。通过使用调整大小后的掩码混合两个结果，生成后续潜扩散步骤的潜变量，即 $z_{\prime\prime t} \odot m_{\text{潜变量}} + z_t \odot (1 - m_{\text{潜变量}})$。如图2所示，其中 $z_{\prime\prime}$ 表示在反向过程中经过增强方向引导的潜在扩散所生成的结果。而 z_{nd} 则表示在前向过程中将相应噪声叠加到输入图像上的结果。在每个去噪步骤中，整个潜变量都会被修改，但后续的混合操作会确保 $m_{\text{潜变量}}$ 之外的部分保持不变。在此阶段，将掩码之外的整个区域替换为输入图像中的对应区域，从而严格保留背景。后续的潜变量去噪过程可确保结果连贯，即使混合后的潜变量并非总是连贯。潜扩散过程完成后，我们使用解码器 $\mathcal{D}(z)$ 将生成的潜变量解码为输出图像，如算法1所示。

此外，还应用了非编辑区域保留损失 L_{NERP}，以引导掩码外的扩散过程，使周围区域趋向输入图像：

$$\mathcal{L}_{NERP}\left(x_0, \hat{x}_t, m\right) = d\left(x_0 \odot (1 - m), \hat{x}_t \odot (1 - m)\right), \tag{8}$$

$$d\left(a, b\right) = \lambda_1\left(LPIPS\left(a, b\right)\right) + \lambda_2\left(MSE\left(a, b\right)\right), \tag{9}$$

$$a = x_0 \odot (1 - m), b = \hat{x}_t \odot (1 - m). \tag{10}$$

其中，$LPIPS$ 是学习感知图像块相似度度量，MSE 是图像之间逐像素差异的 L_2 范数。λ_1 和 λ_1 均设为 0.5。

4 实验

4.1 实现细节

对于扩散模型，我们使用了一个预训练的潜扩散模型 [41]，分辨率为 256×256，包含 1.45 亿个参数，使用 LAION-400M [44] 数据库进行训练。对于 CLIP模型，我们使用了 OpenAI 发布的 ViT-L/14 视觉 Transformer [38]。RDM 的输出大小为 256×256。采样时，我们分别将 λ_1、λ_2 和 CLIP 引导尺度设为 0.5、0.5 和 150。为确保结果质量并保持参数一致性，本研究实验中使用的扩散步数和时间步长均设为 100。使用 RDM 在单张 GeForce RTX 3090 GPU 上生成一张 256×256 图像需要三秒，这与潜在扩散（3 秒）相当，并快于大多数扩散模型（GLIDE、混合扩散和 CLIP引导扩散分别需要 15 秒、27 秒和 3 分钟）。

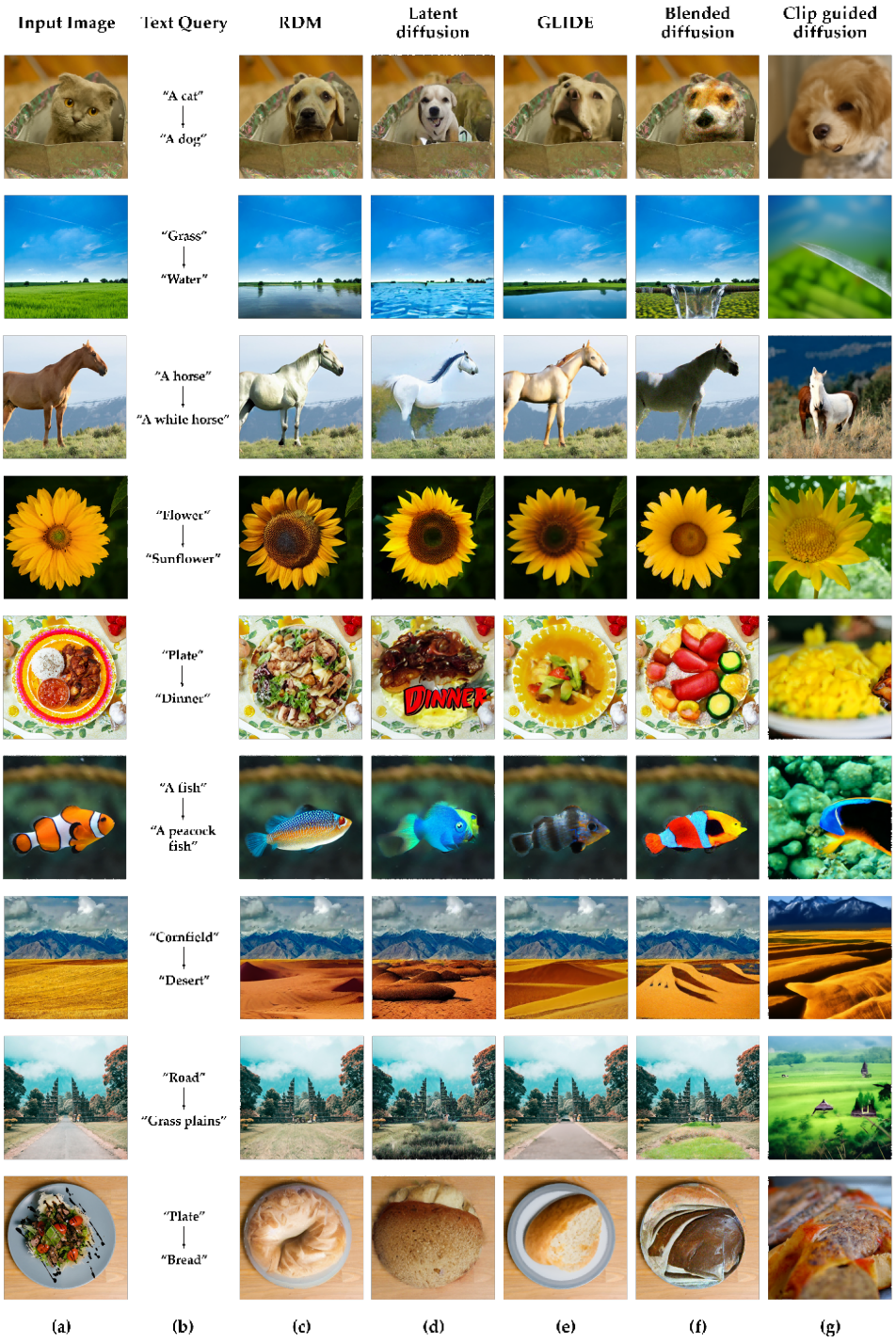


图5. 与最先进技术进行比较，包括潜在扩散、GLIDE、混合扩散和 CLIP引导扩散。

算法1给定文本引导的混合扩散采样
潜扩散模型（$\mu_\theta(z_t)$ $\Sigma_\theta(z_t)$）

要求： 输入图像 x_0、文本引导 t_2、梯度尺度 s、扩散步数 T。
确保： 根据文本引导 t_2生成图像 x_0 。

$t = T$
z_0 =E(x_0)
$z_t \leftarrow$ 从 N (0, I) 中采样
$\hat{z}_T = z_T$
重复
  $t - 1 \leftarrow t$
  $\mu, \Sigma \leftarrow \mu_\theta(\hat{}z_t)$ $\Sigma_\theta(\hat{}z_t)$
  $\hat{z}'_{t-1}$ = 去噪 ($\hat{}z_t, t_2, t$)
  $\hat{x}'_{t-1}$ = D($\hat{z}'_{t-1}$)
    $\hat{}\epsilon_\theta(\hat{}x_{t-1} \mid t_2) \leftarrow (1-s) \cdot \epsilon_\theta(\hat{}x_{t-1} \mid \emptyset) + s \cdot \epsilon_\theta(\hat{x}'_{t-1} \mid t_2)$
  L $\leftarrow$L$_{CLIP}(\hat{x}'_{t-1}, t_2, m)$ +L$_{NERP}(x_0, \hat{x}'_{t-1}, m)$
  $\hat{z}''_{t-1} \leftarrow$ 从 N ($\mu + s\Sigma\nabla_{\hat{x}_{t-1}}$L) 中采样
  $\hat{z}_{t-1} = \hat{z}''_{t-1} \odot m_{\text{潜变量}} + z_{t-1} \odot (1 - m_{\text{潜变量}})$
直到 $t < 0$
$\hat{}x_0$ = D($\hat{}z_0$)



图6. 掩码阈值对操控结果的影响。

4.2 定性评估

我们在多种真实世界图像和编辑文本上测试了我们的方法。这些图像来自网络，包含多种物体类别，包括动物、食物、风景等。输入和结果示例见图 1、3 和4。图 3 展示了我们的方法在t_1 引导下为任意图像生成的分割掩码。结果表明，使用不同图像和文本提示作为输入时，我们的方法不会产生明显的人工编辑痕迹。如图 1中的花朵所示，花朵与背景之间的过渡自然，生成的花瓣具有自然纹理。对于食物操控，可以清楚地看到每一块食物。在动物编辑中，即使待编辑的动物有多种复杂姿势，RDM 也能很好地合成新的动物。如图 4所示，对于相同的图像和文本输入，RDM 能够生成多种不同的结果。总之，我们的方法成功地以高质量和多样化的方式，将文本控制应用于图像实体内容的编辑。

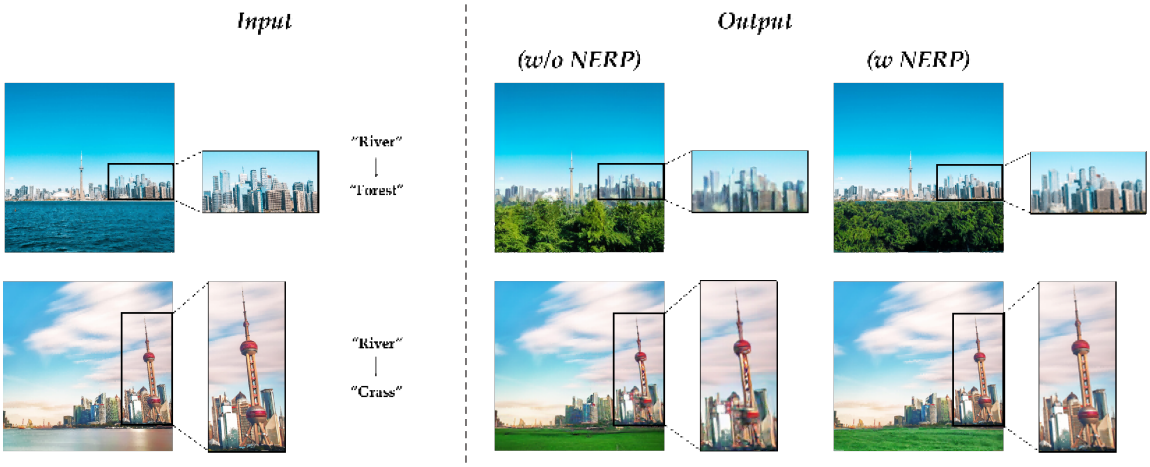


图7. 我们的 RMD 在有无非编辑区域保留组件情况下的定性比较。

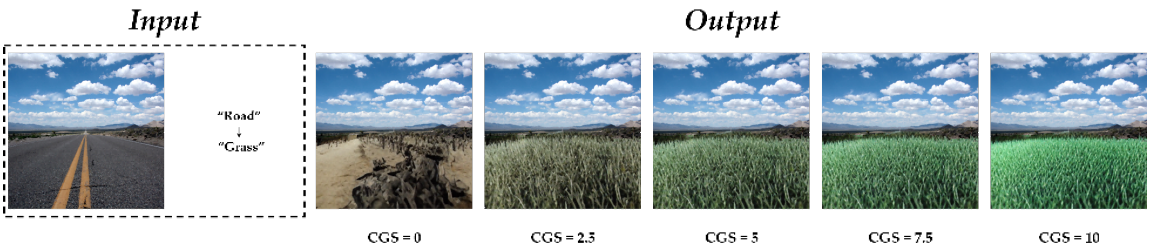


图8. 无分类器引导尺度（CGS）对生成结果的影响。

4.3 与最先进方法的比较

本节将 RDM 与最先进的文本驱动图像编辑方法进行比较，包括潜在扩散 [40],GLIDE [36], 混合扩散 [2] 和 CLIP引导扩散 [10]。图 5 展示了与真实世界图像基线方法的比较。我们的方法与这些方法的主要区别如下。潜在扩散 [40],GLIDE [36],和混合扩散 [2] 要求用户提供需要编辑区域的掩码。它们的非关注区域不参与扩散，因此不存在内容保留问题。CLIP引导扩散 [10] 则无法针对图像的局部区域进行修改。

在此比较中，这些方法所需的掩码由 RDM 的跨模态实体级校准组件生成。从潜在扩散的结果（图 5 的列 (d)）可以看出，[41],即使提供了严格的编辑掩码，生成的新内容也与掩码区域不匹配，且总是小于掩码区域。GLIDE [36] 的结果（图 5 的列 (e)）缺乏细节；例如，向日葵的花瓣非常光滑；马的皮肤没有纹理，马背上的毛发也不够丰富；“草原”的生成失败，GLIDE 也未能很好地理解该文本提示的内容。混合扩散 [2] 生成的图像（图 5 的列 (f)）缺乏真实感，显得很不自然。

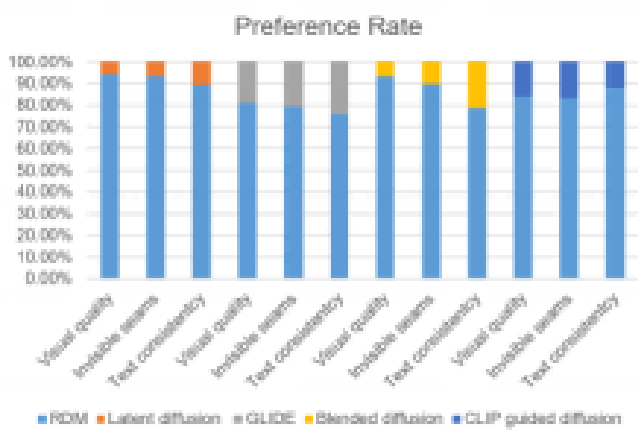


图9. 用户研究结果。每个数字表示与我们的结果相比，其他模型结果获得的投票百分比。

CLIP引导扩散的结果（图5的列 (g)） [10]无法保留图像中未修改区域的内容，并且容易产生过度暗角。

表1. 图像操控的定量比较。我们计算平均 CLIP 分数、SFID分数和图像协调（IH）分数，以衡量视觉质量、文本一致性和图像协调。最佳结果以**粗体** 突出显示，次佳结果则以下划线标示

	RDM	潜在扩散	GLIDE	混合扩散	CLIP 引导的扩散
CLIP 分数 ↑	**0.849**	0.824	0.845	0.822	0.843
SFID分数 ↓	6.54	9.29	**5.88**	17.37	23.42
IH 分数 ↓	**20.7**	22.0	21.8	23.1	/

4.4 定量评估

CLIP 分数。为评估文本描述与修改后的图像之间的语义对齐情况，我们计算 CLIP 分数，即通过 CLIP 编码器获取二者嵌入后计算的余弦相似度。由于我们在推理过程中使用 ViT-B/14 CLIP 模型，为了公平比较，我们使用 ViT-B/32 CLIP [38] 模型计算 CLIP 分数。较高的 CLIP 分数表明输入文本与修改后的图像在语义上相互对齐。结果（表 1^{st} 1 的行）表明，RDM 优于基线模型，这说明其文本图像一致性更佳。

SFID。为评估操纵后的图像质量，我们采用了 SFID [24], ，这是一种简化的 FID，可避免样本特征分布数量有限所导致的数值不稳定性。我们计算了不同方法之间的 SFID 分数，以得出 2^{nd} 表1中的一行。直观而言，SFID 分数越低，COCO [29] 数据集上操纵后的图像质量越高。结果表明，GLIDE 的 SFID 分数最高，RDM 则位居第二。

表2. 定量 ita比较了我们的 RMD 在有无 NERP 组件时的表现。

	含 NERP	不含 NERP
学习感知图像块相似度 ↓	0.039	0.143

协调度得分。为评估编辑区域与未编辑区域之间的协调程度，我们采用 DoveNet [8] 作为定量评估方法。我们使用 DoveNet 生成协调后的图像，并计算协调后图像与操纵后的图像之间的峰值信噪比。图像协调（IH）得分见 3^{rd} 表1中的一行。IH 得分越低，编辑区域与未编辑区域的同步程度越高。我们的方法显著优于基线模型，取得了最高的协调度得分。将掩码融入扩散过程后，编辑区域与未编辑区域之间的一致性表现也有所提升。

用户研究。 接下来，我们进行了用户感知评估。参与者需要选择：哪张图像质量更高、哪张图像看起来更协调且可见的编辑痕迹（接缝）更少，以及哪种图像编辑结果更符合文本内容。我们收集了 70 名参与者针对 36 组图文组合做出的判断，共收集到 2520 张选票。每次比较时，均未透露图像由哪种方法生成。我们在图 9 中报告了支持 RDM 模型的投票百分比。结果表明，我们的方法能够生成更受人们青睐的图像编辑结果。

4.5 消融研究

跨模态实体级校准组件的影响。 为了研究跨模态实体级校准组件对生成图像质量和语义内容的影响，我们在不同阈值 K下测试了图像编辑效果。图 6 显示，当 K 的值较小时，分割出的可编辑场景也较小，其面积在不同程度上小于输入图像中车辆所占的区域。生成车辆的朝向与输入图像中车辆的朝向不一致，即语义一致性存在偏差。随着 K 的值增大，可编辑区域随之扩大，车辆的朝向也趋于一致；但超过一定值后，图像编辑结果便不会发生明显变化。

非编辑区域保留组件的效果。我们对采用和未采用非编辑区域保留（NERP）组件时的编辑结果进行了定性和定量比较。如图7所示，未采用 NERP 组件时，图像中的建筑物会出现更严重的扭曲，并产生异常伪影和纹理，导致整体图像质量下降。学习感知图像块相似度（LPIPS） [48]常用于衡量两幅图像之间的差异。我们使用 LPIPS 衡量编辑前后非关注区域的保留程度。LPIPS 值越低，内容保留效果越好。如表2所示，该组件有助于保留输入图像中非关注区域的图像内容。因此，NERP 组件显著改善了编辑区域之外图像内容的异常扭曲和损坏问题。

无分类器引导尺度的影响。无分类器引导是一种能够显著提升文本条件扩散模型生成质量和图像文本对齐效果的技术。我们研究了无分类器引导对图8中结果的影响。我们发现，不使用无分类器引导得到的结果缺乏竞争力。当 $s > 2.5$ 时，生成的结果更好。在我们的实验中，我们使用默认引导值 $s = 5$。



图 10. 给出了失败案例的示例和对应的源图像。

4.6 失败案例。

通过一些实验，我们观察到 CLIP 对不同编辑器的某些特定解法有明显偏好。如图10所示，给定一张装有咖啡的杯子的图片，我们希望将图像中的“咖啡”编辑为“水”。结果显示，杯中的液体成功变成了水。然而，“水”这一文本与透明杯子密切相关，因此杯子也可能变成玻璃杯。如图10 的第二列所示，空中注入杯中的咖啡未能成功变成水。最后，我们还注意到，当编辑目标的形状与源对象差异很大时，编辑结果并不理想。如图 9 的第三列所示，将“A fish”变为“A jellyfish”产生了额外的边界伪影。需要指出的是，与修改原实体的属性相比，我们的方法更适合生成一个新的实体。

5 结论与未来工作

本文首次探讨了一种新的问题设定：根据任意文本，对图像中特定实体的内容进行编辑。解决这一任务需要控制编辑的位置、编辑后内容的质量与保真度、未编辑内容的保真度，以及文本指导与图像操控之间的一致性等。为了解决这些问题，我们提出了一种用于实体级图像操控的新框架：具备语义对齐和生成能力的区域感知扩散模型。我们为用户提供了一种新工具，用户只需用文本描述需求即可修改图像。未来，我们希望拓展 RDM 的应用，例如更灵活地控制生成区域的位置、形状和大小。

参考文献

[1] Omri Avrahami、Ohad Fried 和 Dani Lischinski。2022。《混合潜在扩散》。*arXiv:2206.02779*（2022）。https://doi.org/10.48550/arXiv.2206.02779[2] Omri Avrahami、Dani Lischinski 和 Ohad Fried。2022。《用于文本驱动自然图像编辑的混合扩散》。载于*2022 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。IEEE。https://doi.org/10.1109/cvpr52688.2022.01767[3] Omer Bar-Tal、Dolev Ofri-Amar、Rafail Fridman、Yoni Kasten 和 Tali Dekel。2022。《Text2live：文本驱动的分层图像与视频编辑》。载于欧洲计算机视觉会议。Springer，707–723。https://doi.org/10.1007/978-3-031-19784-0_41

[4] Dmitry Baranchuk、Ivan Rubachev、Andrey Voynov、Valentin Khrulkov 和 Artem Babenko。2021。使用扩散模型进行标签高效的语义分割。arXiv:2112.03126（2021）。http://arxiv.org/abs/2112.03126v1[5] David Bau、Alex Andonian、Audrey Cui、YeonHwan Park、Ali Jahanian、Aude Oliva 和 Antonio Torralba。2021。按词绘画。arXiv:2103.10951（2021）。https://doi.org/10.48550/arXiv.2103.10951[6] Andrew Brock、Jeff Donahue 和 Karen Simonyan。2018。用于高保真自然图像合成的大规模生成对抗网络训练。arXiv:1809.11096（2018）。http://arxiv.org/abs/1809.11096[7] Jooyoung Choi、Sungwon Kim、Yonghyun Jeong、Youngjune Gwon 和 Sungroh Yoon。2021。ILVR：去噪扩散概率模型的条件控制方法。载于 2021 年 IEEE/CVF国际计算机视觉会议（ICCV）。IEEE。https://doi.org/10.1109/iccv48922.2021.01410[8] Wenyan Cong、Jianfu Zhang、Li Niu、Liu Liu、Zhixin Ling、Weiyuan Li 和 Liqing Zhang。2020。DoveNet：通过域验证实现深度图像协调。载于 2020 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr42600.2020.00842[9] Guillaume Couairon、Asya Grechka、Jakob Verbeek、Holger Schwenk 和 Matthieu Cord。2022。FlexIT：迈向灵活的语义图像翻译。载于 2022 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr52688.2022.01773[10] Katherine Crowson。2022。CLIP 引导的扩散 HQ 256x256。https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj[11] Prafulla Dhariwal 和 Alex Nichol。2021。扩散模型在图像合成方面胜过生成对抗网络。载于《神经信息处理系统进展》。NeurIPS，8780–8794。https://doi.org/10.48550/arXiv.2105.05233[12] Hao Dong、Simiao Yu、Chao Wu 和 Yike Guo。2017。通过对抗学习实现语义图像合成。载于 2017 年 IEEE国际计算机视觉会议（ICCV）。IEEE。https://doi.org/10.1109/iccv.2017.608[13] Vincent Dumoulin、Ethan Perez、Nathan Schucher、Florian Strub、Harm Vries、Aaron Courville 和 Yoshua Bengio。2018。逐特征变换。《Distill》3，7（2018 年 7 月）。https://doi.org/10.23915/distill.00011[14] Patrick Esser、Robin Rombach 和 Bjorn Ommer。2021。驾驭 Transformer 进行高分辨率图像合成。载于 2021 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr46437.2021.01268[15] Rinon Gal、Or Patashnik、Haggai Maron、Amit H. Bermano、Gal Chechik 和 Daniel Cohen-Or。2022。StyleGAN-NADA：由 CLIP 引导的图像生成器域适应。《ACM图形学汇刊》41，4（2022 年 7 月），1–13。https://doi.org/10.1145/3528223.3530164[16] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020。去噪扩散概率模型。《神经信息处理系统进展》（NeurIPS）（2020），6840–6851。https://doi.org/10.48550/arXiv.2006.11239[17] Jonathan Ho 和 Tim Salimans。2021。无分类器扩散引导。载于 NeurIPS 2021 深度生成模型与下游应用研讨会。https://doi.org/10.48550/arXiv.2207.12598[18] Nisha Huang、Fan Tang、Weiming Dong 和 Changsheng Xu。2022。绘出你的艺术之梦：通过多模态引导扩散实现多样化数字艺术合成。载于第 30 届 ACM国际多媒体会议论文集。ACM。https://doi.org/10.1145/3503161.3548282[19] Xun Huang 和 Serge Belongie。2017。通过自适应实例归一化实时实现任意风格迁移。载于 2017 年 IEEE国际计算机视觉会议（ICCV）。IEEE。https://doi.org/10.1109/iccv.2017.167[20] Wentao Jiang、Ning Xu、Jiayun Wang、Chen Gao、Jing Shi、Zhe Lin 和 Si Liu。2021。通过跨模态循环机制实现语言引导的全局图像编辑。载于 2021 年 IEEE/CVF国际计算机视觉会议（ICCV）。IEEE。https://doi.org/10.1109/iccv48922.2021.00212[21] Alexia Jolicoeur-Martineau、Rémi Piché-Taillefer、Ioannis Mitliagkas 和 Remi Tachet des Combes。2021。用于图像生成的对抗性评分匹配与改进采样。载于国际学习表征会议。ICLR。https://doi.org/10.48550/arXiv.2009.05475[22] Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。2020。分析并提升 StyleGAN 的图像质量。载于 2020 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr42600.2020.00813[23] Bahjat Kawar、Michael Elad、Stefano Ermon 和 Jiaming Song。2022。去噪扩散复原模型。arXiv:2201.11793（2022）。http: //arxiv.org/abs/2201.11793v1[24] Chung-Il Kim、Meejoung Kim、Seungwon Jung 和 Eenjun Hwang。2020。生成对抗网络的简化 Fréchet距离。《Sensors》20，6（2020 年 3 月），1548。https://doi.org/10.3390/s20061548[25] Gwanghyun Kim、Taesung Kwon 和 Jong Chul Ye。2022。DiffusionCLIP：用于鲁棒图像操控的文本引导扩散模型。载于 2022 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr52688.2022.00246[26] Diederik P Kingma 和 Max Welling。2013。自动编码变分贝叶斯。arXiv:1312.6114（2013）。http://arxiv.org/abs/1312.6114[27] Gihyun Kwon 和 Jong Chul Ye。2022。CLIPstyler：通过单一文本条件进行图像风格迁移。载于 2022 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr52688.2022.01753[28] Bowen Li、Xiaojuan Qi、Thomas Lukasiewicz 和 Philip H.S. Torr。2020。ManiGAN：文本引导的图像操控。载于 2020 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr42600.2020.00790[29] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C. Lawrence Zitnick。2014。Microsoft COCO：上下文中的常见物体。载于《计算机视觉——ECCV 2014》。Springer International Publishing，740–755。https://doi.org/10.1007/978-3- 319-10602-1 48_[30] Xihui Liu、Zhe Lin、Jianming Zhang、Handong Zhao、Quan Tran、Xiaogang Wang 和 Hongsheng Li。2020。Open-Edit：基于开放词汇指令的开放域图像操控。载于《计算机视觉——ECCV 2020》。Springer International Publishing，89–106。https://doi.org/10.1007/978-3-030-58621-8 6_[31] Xihui Liu、Dong Huk Park、Samaneh Azadi、Gong Zhang、Arman Chopikyan、Yuxiao Hu、Humphrey Shi、Anna Rohrbach 和 Trevor Darrell。2023。更多控制，免费获得！通过语义扩散引导实现图像合成。（2023 年 1 月）。https://doi.org/10.1109/wacv56688.2023.00037 提交至 ACM 的稿件

[32] Timo Luddecke 和 Alexander Ecker。2022。使用文本和图像提示进行图像分割。载于 2022 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr52688.2022.00695[33] Elman Mansimov、Emilio Parisotto、Jimmy Lei Ba 和 Ruslan Salakhutdinov。2015。利用注意力从图像描述生成图像。arXiv:1511.02793（2015）。http://arxiv.org/abs/1511.02793[34] Sangwoo Mo、Minsu Cho 和 Jinwoo Shin。2019。Instagan：实例感知图像到图像转换。载于国际学习表征会议。ICLR。https://openreview.net/forum?id=ryxwJhC9YX[35] Seonghyeon Nam、Yunji Kim 和 Seon Joo Kim。2018。文本自适应生成对抗网络：使用自然语言操控图像。神经信息处理系统进展（NeurIPS）（2018），42–51。https://doi.org/10.48550/arXiv.1810.11919[36] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2021。GLIDE：借助文本引导扩散模型实现照片级真实感图像生成与编辑。arXiv:2112.10741（2021）。http://arxiv.org/abs/2112.10741v3[37] Or Patashnik、Zongze Wu、Eli Shechtman、Daniel Cohen-Or 和 Dani Lischinski。2021。StyleCLIP：StyleGAN 图像的文本驱动操控。载于 2021 年 IEEE/CVF国际计算机视觉会议（ICCV）。IEEE。https://doi.org/10.1109/iccv48922.2021.00209[38] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。利用自然语言监督学习可迁移视觉模型。载于国际机器学习会议（ICML）。PMLR，8748–8763。https://doi.org/10.48550/arXiv.2103.00020[39] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022。利用 CLIP 潜变量进行分层文本条件图像生成。arXiv:2204.06125（2022）。http://arxiv.org/abs/2204.06125[40] Robin Rombach。2022。潜在扩散 LAION 400M 模型：文本到图像修复。https://colab.research.google.com/github/loboere/latent- diffusion-inpainting-colab/blob/main/Latent_Diffusion_LAION_400M_model_text_to_image_inpainting.ipynb#scrollTo=NUmmV5ZvrPbP[41] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 BjornOmmer。2022。利用潜在扩散模型进行高分辨率图像合成。载于 2022 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr52688.2022.01042[42] Olaf Ronneberger。2017。特邀报告：用于生物医学图像分割的 U-Net 卷积网络。载于 Informatik aktuell。Springer Berlin Heidelberg，3–3。https://doi.org/10.1007/978-3-662-54345-0_3[43] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。2022。具有深度语言理解能力的照片级真实感文本到图像扩散模型。arXiv:2205.11487（2022）。https://doi.org/10.48550/arXiv.2205.11487[44] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。2021。LAION-400M：由 CLIP 筛选的 4 亿图文对开放数据集。arXiv:2111.02114（2021）。http://arxiv.org/abs/2111.02114[45] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。2015。利用非平衡热力学进行深度无监督学习。载于国际机器学习会议。PMLR，ICML，2256–2265。https://doi.org/10.48550/arXiv.1503.03585[46] Jiaming Song、Chenlin Meng 和 Stefano Ermon。2021。去噪扩散隐式模型。载于国际学习表征会议。ICLR。https://doi.org/10.48550/arXiv.2010.02502[47] Jianan Wang、Guansong Lu、Hang Xu、Zhenguo Li、Chunjing Xu 和 Yanwei Fu。2022。ManiTrans：通过逐词元语义对齐与生成实现实体级文本引导图像操控。载于 2022 年 IEEE/CVF计算机视觉与模式识别会议（CVPR）。IEEE。https://doi.org/10.1109/cvpr52688.2022.01044[48] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。2018。深层特征作为感知度量的惊人效用。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）。586–595。https://doi.org/10.48550/arXiv.1801.03924