

基于扩散模型的图像编辑：一项调查

是

黄毅*、黄建城*、刘一凡*、严明福*、吕嘉熙*、刘建庄*、*Senior Member, IEEE*、熊伟、张赫、陈世峰、曹亮亮, *Senior Member, IEEE*

摘要—去噪扩散模型已成为各种图像生成和编辑任务的强大工具，以无条件或输入条件的方式促进视觉内容的合成。它们背后的核心思想是学习逆转逐渐向图像添加噪声的过程，从而使它们能够从复杂的分布中生成高质量的样本。在本调查中，我们详尽概述了使用扩散模型进行图像编辑的现有方法，涵盖了该领域的理论和实践方面。我们从多个角度深入分析和分类这些作品，包括学习策略、用户输入条件以及可以完成的一系列特定编辑任务。此外，我们特别关注图像修复和修复，并探索早期的传统上下文驱动方法和当前的多模态条件方法，对它们的方法进行全面分析。为了进一步评估文本引导的图像编辑算法的性能，我们提出了一个系统基准 EditEval，它具有创新的指标 LMM 分数。最后，我们解决了当前的局限性并展望了未来研究的一些潜在方向。随附的存储库发布于 <https://github.com/SiatMLab/Awesome-Diffusion-Model-Based-Image-Editing-Methods>。

索引词——扩散模型、图像编辑、AIGC

1 引言

在人工智能生成内容 (AIGC) 领域，人工智能被用来创建和修改数字内容，而图像编辑被视为一个重要的创新和实际应用领域。与通过最少的输入创建新图像的图像生成不同，图像编辑涉及改变图像的外观、结构或内容，涵盖从细微调整到重大转变的一系列变化。这类研究是数字媒体、广告和科学研究等各个领域的基础，在这些领域，改变视觉内容至关重要。图像编辑的发展反映了数字技术的进步，从手动、劳动密集型的过程发展到由基于学习的算法驱动的先进数字技术。这一发展中的一个关键进步是生成对抗网络 (GAN) [1] – [6] 的引入，它大大增强了创造性图像处理的可能性。

最近，AIGC [1]、[7] – [15] 中出现了扩散模型，这在视觉生成任务中取得了显著突破。扩散模型受到非平衡热力学原理的启发 [15]，其工作原理是逐渐向数据中添加噪声，然后学习从随机噪声中逆转此过程，直到生成与源数据分布匹配的所需数据。它们大致可分为基于去噪扩散的 [15] – [18] 和基于得分匹配的 [19] – [23]。它们的适应性和有效性已在各种领域得到广泛应用。

- Y. Huang, J. Huang, M. Yan, and J. Lv are with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, and also with University of Chinese Academy of Sciences, Beijing, China. (E-mail: yi.huang@siat.ac.cn)
- Y. Liu is with Southern University of Science and Technology, Shenzhen, China, and also with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (E-mail: jz.liu@siat.ac.cn, shifeng.chen@siat.ac.cn)
- J. Liu and S. Chen are with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (E-mail: jz.liu@siat.ac.cn, shifeng.chen@siat.ac.cn)
- W. Xiong and H. Zhang are with Adobe Inc, San Jose, USA.
- L. Cao is with Apple Inc, Cupertino, USA.
- * denotes equal contributions. S. Chen is the corresponding author.

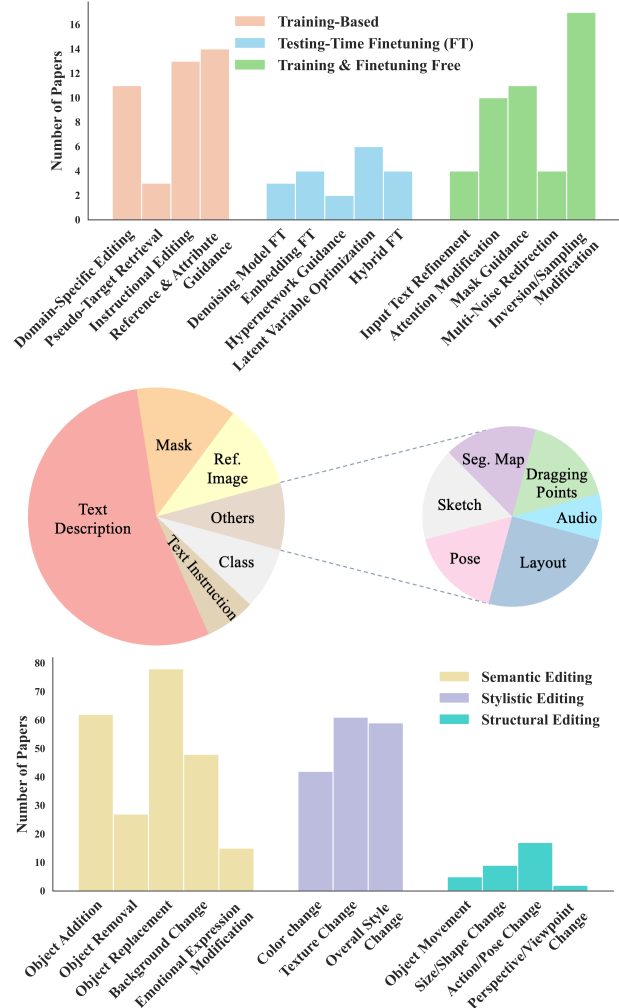


图 1：基于扩散模型的图像编辑研究出版物的统计概览。顶部：学习策略。中间：输入条件。底部：编辑任务。

例如图像生成[24]–[38]、视频生成[39]–[56]、图像恢复[57]–[71]和图像编辑等任务。

近年来，该领域的大量研究出版物表明，人们对扩散模型在图像编辑中的应用兴趣激增。与以前的研究相比，这种日益增长的关注凸显了扩散模型在提高图像编辑性能方面的潜力和多功能性。鉴于这一重大进步，系统地回顾和总结这些贡献至关重要。然而，现有的关于扩散模型的调查文献集中于其他特定的视觉任务[72]–[75]，如视频应用[73]或图像恢复和增强[74]，[75]。一些提到图像编辑的调查往往只提供粗略的概述[76]–[83]，缺少对方法的详细和有针对性的探索。

为了弥补这一差距，我们进行了一项调查，以提供深入而全面的分析，该分析明确侧重于图像编辑。我们深入研究了该领域的方法、输入条件和扩散模型实现的广泛编辑任务。该调查严格审查了 100 多篇研究论文，根据学习策略将它们分为三个主要类别：*training-based approaches*、*testing-time fine-tuning approaches* 和 *training and finetuning free approaches*。每个类根据其核心技术进一步划分，并分别在第 4、5 和 6 节中介绍详细讨论。我们还探讨了这些方法中使用的 10 种不同类型的输入条件，包括文本、蒙版、参考 (Ref.) 图像、类、布局、姿势、草图、分割 (Seg.) 图、音频和拖动点，以展示扩散模型在不同图像编辑场景中的适应性。此外，我们的调查将图像编辑任务分为三大类，*semantic editing*、*stylistic editing* 和 *structural editing*，涵盖 12 种特定类型。图 1 直观地表示了 *learning strategies*、*input conditions* 和 *editing task categories* 中研究的统计分布。此外，我们特别关注修复和覆盖修复，它们共同构成了一种独特的编辑类型。我们探讨了早期的传统和当前的多模态条件方法，并在第 7 节中对它们的方法进行了全面的分析。我们还介绍了 EditEval，这是一个旨在评估文本引导图像编辑算法的基准，详见第 8 节。特别是，通过利用大型多模态模型 (LMM) 的高级视觉语言理解能力，提出了一种有效的评估指标 LMM 分数。最后，我们在第 9 节中介绍了一些当前的挑战和潜在的未来趋势作为展望。

总之，本综述旨在系统地对基于扩散模型的图像编辑的大量研究进行分类和批判性评估。我们的目标是提供全面的资源，不仅综合当前的研究成果，而且还能指导这一快速发展的领域的未来研究方向。

2 背景

2.1 扩散模型

扩散模型对生成式人工智能领域产生了深远的影响，催生了大量属于该范畴的方法。本质上，这些模型基于一项称为扩散的关键原理，该原理逐渐向某些分布的数据样本添加噪声，将其转换为预定义的典型简单分布（例如高斯分布），然后迭代地逆向执行此过程以生成与原始数据匹配的数据

分布。扩散模型与早期生成模型的区别在于，扩散模型可以在迭代时间步骤中动态执行，涵盖时间上的向前和向后运动。

对于每个时间步长 t ，噪声潜变量 $\mathbf{z}_t$ 描绘了当前状态。时间步长 t 在正向扩散过程中逐渐增加，在反向扩散过程中趋向于 0。值得注意的是，文献中没有对正向扩散的 $\mathbf{z}_t$ 和反向扩散的 $\mathbf{z}_t$ 进行明显区分。在正向扩散中设 $\mathbf{z}_t \sim q(\mathbf{z}_t | \mathbf{z}_{t-1})$ ，在反向扩散中设 $\mathbf{z}_{t-1} \sim p(\mathbf{z}_{t-1} | \mathbf{z}_t)$ 。在此，我们将 T 表示为 $0 < t \leq T$ ，即有限情况下的最大时间步长。 $t = 0$ 处的初始数据分布用 $\mathbf{z}_0 \sim q(\mathbf{z}_0)$ 表示，并慢慢被加性噪声污染到 $\mathbf{z}_0$ 。扩散模型通过参数化模型 $p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t)$ 在反时间方向上逐渐消除噪声。该模型近似于理想的（尽管无法实现的）去噪分布 $p(\mathbf{z}_{t-1} | \mathbf{z}_t)$ 。降噪扩散概率模型 (DDPM)，如 Ho *et al.* [16] 中所述，有效地利用马尔可夫链来促进有限时间步骤中的前向和后向过程。前向扩散过程。此过程用于将数据分布转换为预定义分布，例如高斯分布。转换表示为：

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t | \sqrt{1 - \beta_t} \mathbf{z}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

其中超参数集 $0 < \beta_{1:T} < 1$ 表示在每个连续步骤中引入的噪声方差。此扩散过程可以通过单步方程简要表示：

$$q(\mathbf{z}_t | \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t | \sqrt{\alpha_t} \mathbf{z}_0, (1 - \alpha_t) \mathbf{I}), \quad (2)$$

其中 $\alpha_t = 1 - \beta_t$ 和 $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ ，如 Sohl-Dickstein *et al.* [15] 中所述。因此，无需考虑中间时间步骤， $\mathbf{z}_t$ 可直接通过以下方式采样：

$$\mathbf{z}_t = \sqrt{\alpha_t} \cdot \mathbf{z}_0 + \sqrt{1 - \alpha_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (3)$$

逆扩散过程。这里的主要目标是学习正向扩散过程的逆过程，目的是生成与原始未改变的数据样本 $\mathbf{z}_0$ 紧密一致的分布。在图像编辑的背景下， $\mathbf{z}_0$ 表示编辑后的图像。实际上，这是通过使用 UNet 架构来学习 p 的参数化版本来实现的。假设正向扩散过程近似于 $q(\mathbf{z}_T) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$ ，可学习转换的公式表示为：

$$p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1} | \mu_\theta(\mathbf{z}_t, \bar{\alpha}_t), \Sigma_\theta(\mathbf{z}_t, \bar{\alpha}_t)), \quad (4)$$

其中函数 μ_θ 和 Σ_θ 是可学习的参数。此外，对于条件公式 $p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t, \mathbf{c})$ ，其以图像编辑中的外部变量 $\mathbf{c}$ （为条件， $\mathbf{c}$ 可以是源图像），模型变为 $\mu_\theta(\mathbf{z}_t, \mathbf{c}, \bar{\alpha}_t)$ 和 $\Sigma_\theta(\mathbf{z}_t, \mathbf{c}, \bar{\alpha}_t)$ 。

优化。在学习前向过程时，引导反向扩散的优化策略涉及最小化正向和反向序列联合分布之间的 Kullback-Leibler (KL) 散度。这些数学定义为：

$$p_\theta(\mathbf{z}_0, \dots, \mathbf{z}_T) = p(\mathbf{z}_T) \prod_{t=1}^T p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t), \quad (5)$$

$$q(\mathbf{z}_0, \dots, \mathbf{z}_T) = q(\mathbf{z}_0) \prod_{t=1}^T q(\mathbf{z}_t | \mathbf{z}_{t-1}), \quad (6)$$

leading to the minimization of:

$$\begin{aligned} \text{KL}(q(\mathbf{z}_0, \dots, \mathbf{z}_T) \| p_\theta(\mathbf{z}_0, \dots, \mathbf{z}_T)) \\ \geq \mathbb{E}_{q(\mathbf{z}_0)} [-\log p_\theta(\mathbf{z}_0)] + c, \end{aligned} \quad (7)$$

这在 Ho *et al.* [16] 中有详细说明, 常数 c 与优化 θ 无关。等式 7 中的 KL 散度表示数据对数似然的变分界限 ($\log p_\theta(\mathbf{z}_0)$)。此 KL 散度用作损失, 并在 DDPM 中最小化。实际上, Ho *et al.* [16] 采用了此损失的重新加权版本作为更简单的去噪损失:

$$\mathbb{E}_{t \sim \mathcal{U}(1, T), \mathbf{z}_0 \sim q(\mathbf{z}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\lambda(t) \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|^2], \quad (8)$$

其中 $\lambda(t) > 0$ 表示加权函数, $\mathbf{z}_t$ 利用公式 3 得到, ϵ_θ 表示基于 $\mathbf{z}_t$ 和 t 预测噪声 ϵ 的网络。

DDIM 采样和反演。处理真实图像 $\mathbf{z}_0$ 时, 流行的编辑方法 [84]、[85] 首先利用特定的反演方案将 $\mathbf{z}_0$ 反演为相应的 $\mathbf{z}_T$ 。随后, 从 $\mathbf{z}_T$ 开始采样, 采用某种编辑策略来生成编辑结果 $\tilde{\mathbf{z}}_0$ 。在理想情况下, 从 $\mathbf{z}_T$ 直接采样 (不进行任何编辑) 应产生与 $\mathbf{z}_0$ 非常相似的 $\tilde{\mathbf{z}}_0$ 。 $\tilde{\mathbf{z}}_0$ 与 $\mathbf{z}_0$ 的显著偏差 (称为重建失败) 表示编辑后的图像无法保持 $\mathbf{z}_0$ 中未改变区域的完整性。因此, 使用确保 $\tilde{\mathbf{z}}_0 \approx \mathbf{z}_0$ 的反演方法至关重要。

DDIM 采样方程[18]为:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \frac{\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{z}_t, t)}{\sqrt{\bar{\alpha}_t}} + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(\mathbf{z}_t, t), \quad (9)$$

也可以表达为:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \frac{\mathbf{z}_{t-1} - \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(\mathbf{z}_{t-1}, t)}{\sqrt{\bar{\alpha}_{t-1}}} + \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{z}_t, t). \quad (10)$$

尽管方程 10 似乎提供了从 $\mathbf{z}_{t-1}$ 到 $\mathbf{z}_t$ 的理想反演, 但问题在于 $\mathbf{z}_t$ 的未知性质, 它也用作网络 $\epsilon_\theta(\mathbf{z}_t, t)$ 的输入。为了解决这个问题, DDIM 反演 [18] 在 $\mathbf{z}_{t-1} \approx \mathbf{z}_t$ 的假设下运行, 并将方程 10 右侧的 $\mathbf{z}_t$ 替换为 $\mathbf{z}_{t-1}$, 从而得到以下近似值:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \frac{\mathbf{z}_{t-1} - \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(\mathbf{z}_{t-1}, t)}{\sqrt{\bar{\alpha}_{t-1}}} + \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{z}_{t-1}, t). \quad (11)$$

文本条件和无分类器指导。文本条件扩散模型旨在合成由文本提示 P 引导的随机噪声 $\mathbf{z}_T$ 的结果。在采样过程中的推理过程中, 噪声估计网络 $\epsilon_\theta(\mathbf{z}_t, t, C)$ 用于预测噪声 ϵ , 其中 $C = \psi(P)$ 表示文本嵌入。此过程有条不紊地在 T 个步骤中从 $\mathbf{z}_t$ 中去除噪声, 直到获得最终结果 $\mathbf{z}_0$ 。

在文本条件图像生成领域, 确保对生成的输出具有实质性的文本影响和控制至关重要。为此, Ho *et al.* [86] 引入了无分类器指导的概念, 这是一种融合条件预测和非条件预测的技术。更具体地说, 让 $\emptyset = \psi(\text{" "})^1$ 表示空文本嵌入。当与指导尺度 w 结合时, 无分类器指导预测形式化为:

$$\epsilon_\theta(\mathbf{z}_t, t, C, \emptyset) = w \epsilon_\theta(\mathbf{z}_t, t, C) + (1 - w) \epsilon_\theta(\mathbf{z}_t, t, \emptyset). \quad (12)$$

1. 占位符 $\emptyset$ 通常用于负面提示, 以防止某些属性在生成的图像中显现。

在这个公式中, $\epsilon_\theta(\mathbf{z}_t, t, C, \emptyset)$ 取代了采样方程 9 中的 $\epsilon_\theta(\mathbf{z}_t, t)$ 。 w 的值通常在 $[1, 7.5]$ 之间, 如 [26]、[27] 中所述, 它决定了文本控制的程度。 w 值越高, 生成过程中文本驱动的影响就越强。

2.2 相关任务

2.2.1 Conditional Image Generation

虽然我们主要关注的是图像编辑中的扩散模型, 但了解条件图像生成等相关领域也很重要。与涉及更改现有图像某些部分的图像编辑不同, 条件图像生成是在指定条件的指导下从头创建新图像。早期的努力 [31]、[32]、[87]–[90] 通常涉及类条件图像生成, 通常在采样期间通过额外的预训练分类器合并类诱导梯度。然而, Ho *et al.* [86] 引入了无分类器指导, 它不依赖于外部分类器, 并且允许使用更通用的条件 (例如文本) 作为指导。

文本到图像 (T2I) 生成。GLIDE [34] 是第一篇使用文本直接从高维像素级别指导图像生成的论文, 它取代了类条件扩散模型中的标签。类似地, Imagen [27] 使用级联框架在像素空间中更有效地生成高分辨率图像。另一条研究路线首先将图像投影到低维空间, 然后在该潜在空间中应用扩散模型。代表性论文包括稳定扩散 (SD) [26]、VQ 扩散 [91] 和 DALL-E 2 [25]。在这些开创性研究之后, 提出了大量论文 [37]、[92]–[97], 在过去两年中推动了该领域的发展。

附加条件。除了文本之外, 还使用更具体的条件来实现图像合成中更高的保真度和更精确的控制。GLIGEN [98] 在预训练的 T2I 扩散模型的每个块中, 在原始自注意层和交叉注意层之间插入一个门控自注意层, 用于生成以接地框为条件的图像。Make-A-Scene [99] 和 SpaText [100] 使用分割蒙版来指导生成过程。除了分割图之外, ControlNet [101] 还可以合并其他类型的输入, 例如深度图、法线图、精明边缘、姿势和草图作为条件。其他方法, 如 UniControlNet [102]、UniControl [103]、Composer [104] 和 T2I-Adapter [105] 集成了各种条件输入并包含附加层, 从而增强了由这些条件控制的生成过程。

定制图像生成。在条件图像生成中, 与图像编辑密切相关的是创建个性化图像的任务。此任务侧重于生成保持一定身份的图像, 通常由同一主题的几张参考图引导。两种早期通过少量图像解决这种定制生成的方法是文本反转 [106] 和 DreamBooth [107]。具体来说, 文本反转学习一个唯一的标识符词来表示一个新主题, 并将这个词合并到文本编码器的词典中。另一方面, DreamBooth 通过使用一些参考图像微调整个 Imagen [27] 模型, 将一个全新的、罕见的词与特定主题绑定在一起。为了有效地组合多个新概念, CustomDiffusion [108] 仅优化了稳定扩散 [26] 中的交叉注意参数, 表示新概念并对多概念组合进行联合训练。

在这些基础工作之后，后续方法[109] – [117]对生成的图像提供了更精细的控制，提高了输出的精度和准确度。

2.2.2 Image Restoration and Enhancement

图像恢复 (IR) 是低级视觉中的一项关键任务，旨在提高受各种退化污染的图像的质量。扩散模型的最新进展促使研究人员研究其在图像恢复中的潜力。先驱性的努力将扩散模型整合到这项任务中，超越了以前基于 GAN 的方法。

输入图像作为条件。生成模型对各种图像恢复任务做出了重大贡献，例如超分辨率 (SR) 和去模糊 [12]、[13]、[29]、[118]、[119]。通过重复细化的超分辨率 (SR3) [57] 利用 DDPM 通过随机迭代去噪过程进行条件图像生成。级联扩散模型 [31] 按顺序采用多个扩散模型，每个模型都会生成分辨率不断提高的图像。SRDiff [118] 遵循 SR3 的紧密实现。SRDiff 和 SR3 之间的主要区别在于 SR3 直接预测目标图像，而 SRDiff 预测输入和输出图像之间的差异。

非空间空间中的恢复。一些基于扩散模型的 IR 方法专注于其他空间。例如，Refusion [63]、[120] 采用均值回归图像恢复 (IR)-SDE 将目标图像转换为其退化对应图像。它们利用自动编码器将输入图像压缩为其潜在表示，并使用跳过连接来访问多尺度细节。Chen *et al.* [121] 采用类似的方法，提出了一种称为分层集成扩散模型的两阶段策略。从空间域到小波域的转换是无损的，并且具有显著的优势。例如，WaveDM [67] 修改了低频带，而 WSGM [122] 或 ResDiff [60] 相对于低分辨率图像调节高频带。BDCE [123] 在深度曲线空间中设计了一个引导扩散模型，用于高分辨率低光图像增强。

T2I 先验使用。T2I 信息的结合被证明是有利的，因为它允许使用预训练的 T2I 模型。这些模型可以通过添加针对 IR 任务量身定制的特定层或编码器进行微调。Wang *et al.* 使用 StableSR [124] 将这一概念付诸实践。StableSR 的核心是一个时间感知编码器，与冻结的稳定扩散模型 [26] 一起训练。此设置无缝集成了可训练的空间特征变换层，从而能够根据输入图像进行调节。DiffBIR [125] 使用预训练的 T2I 扩散模型进行盲图像恢复，具有两级流水线和可控模块。CoSeR [126] 引入了认知超分辨率，融合了图像外观和语言理解。SUPIR [127] 利用生成先验、模型缩放和 2000 万张图像的数据集，在文本提示的指导下进行高级恢复，具有负质量提示和恢复引导采样方法。

基于投影的方法。这些方法旨在从输入图像中提取固有结构或纹理，以补充每一步生成的图像并确保数据一致性。ILVR [65] 将输入图像中的低频信息投影到输出图像，确保数据一致性并建立改进的条件。为了解决这个问题并增强数据一致性，最近的一些研究 [70]、[71]、[128] 采取了

旨在利用贝叶斯定理来估计后验分布的不同方法。

基于分解的方法。这些方法将 IR 任务视为线性逆问题。去噪扩散恢复模型 (DDRM) [66] 利用预训练的去噪扩散生成模型来解决线性逆问题，展示了在不同测量噪声水平下在超分辨率、去模糊、修复和着色方面的多功能性。去噪扩散零空间模型 (DDNM) [68] 代表了另一种基于分解的零样本方法，适用于图像 SR 以外的广泛线性 IR 问题，例如着色、修复和去模糊。它利用范围零空间分解方法 [129]、[130] 来有效应对各种 IR 挑战。

3 图像编辑的分类

扩散模型除了在图像生成、修复和增强方面取得了重大进展外，在图像编辑方面也取得了显著突破，与之前占主导地位的 GAN 相比，扩散模型具有更强的可控性。与专注于从头创建新图像的图像生成和旨在修复和提高退化图像质量的图像修复和增强不同，图像编辑涉及对现有图像的外观、结构或内容进行修改，包括添加对象、替换背景和更改纹理等任务。

在本调查中，我们根据学习策略将图像编辑论文分为三个主要组：*training-based approaches*、*testing-time finetuning approaches* 和 *training and finetuning free approaches*，分别在第 4、5 和 6 节中进行了详细说明。此外，我们探讨了这些方法用于控制编辑过程的 10 种输入条件，包括文本、蒙版、参考 (Ref.) 图像、类、布局、姿势、草图、分割 (Seg.) 图、音频和拖动点。此外，我们研究了这些方法可以完成的 12 种最常见的编辑类型，这些类型分为以下定义的三大类。

- *Semantic Editing*：此类别包括对图像内容和叙事的更改，影响所描绘场景的故事、背景或主题元素。此类别中的任务包括添加对象 (Obj. Add.)、移除对象 (Obj. Remo.)、替换对象 (Obj. Repl.)、更改背景 (Bg. Chg.) 和修改情绪表达 (Emo. Expr. Mod.)。
- *Stylistic Editing*：此类别侧重于增强或改变图像的视觉风格和美学元素，而不改变其叙事内容。此类别中的任务包括颜色变化 (颜色变化)、纹理变化 (文本变化) 和整体风格变化 (风格变化)，涵盖艺术风格和现实风格。
- *Structural Editing*：此类别涉及图像中元素的空间排列、定位、视点和特征的变化，强调场景中对象的组织和呈现。此类别中的任务包括对象移动 (Obj. Move.)、对象大小和形状变化 (Obj. Size. Chg.)、对象动作和姿势变化 (Obj. Act. Chg.) 以及视角/视点变化 (Persp./View. Chg.)。

表1对所调查论文的多视角分类进行了全面总结，方便快速搜索。

表 1：从多个角度对基于扩散模型的图像编辑方法进行全面分类。这些方法根据训练、测试时微调和训练与微调自由进行颜色呈现。输入条件包括文本、类、参考 (Ref.) 图像、分割 (Seg.) 图、姿势、蒙版、布局、草图、拖动点和音频。根据源论文中提供的实验结果，任务能力（语义、风格和结构）用复选标记 (✓) 标记。

Method	Condition(s)	Semantic Editing					Stylistic Editing			Structural Editing			
		Obj. Add.	Obj. Remo.	Obj. Repl.	Bg. Chg.	Emo. Expr. Mod.	Color Chg.	Style Chg.	Text. Chg.	Obj. Move.	Obj. Size. Chg.	Obj. Act. Chg.	Persp. /View. Chg.
DiffusionCLIP [131]	Text, Class			✓		✓	✓	✓	✓				
Asyrrp [132]	Text, Class			✓		✓			✓				
EffDiff [133]	Class			✓		✓	✓	✓	✓				
DiffStyler [134]	Text, Class							✓					
StyleDiffusion [135]	Ref. Image, Class							✓					
UNIT-DDPM [136]	Class							✓	✓				
CycleNet [137]	Class			✓				✓					
Diffusion Autoencoders [138]	Class			✓		✓			✓				
HDAE [139]	Class			✓		✓			✓				
EGSDE [140]	Class			✓									
Pixel-Guided Diffusion [141]	Seg. Map, Class								✓		✓	✓	
PbE [142]	Ref. Image	✓		✓									
RIC [143]	Ref. Image, Sketch	✓		✓	✓		✓				✓		
ObjectStitch [144]	Ref. Image	✓											
PhD [145]	Ref. Image, Layout	✓		✓				✓					
DreamInpainter [146]	Ref. Image, Mask, Text	✓	✓							✓			✓
Anydoor [147]	Ref. Image, Mask	✓		✓						✓			
FADING [148]	Text								✓				
PAIR Diffusion [149]	Ref. Image, Text	✓		✓			✓		✓				
SmartBrush [150]	Text, Mask	✓		✓	✓								
IIR-Net [151]	Text				✓		✓	✓					
PowerPaint [152]	Text, Mask	✓	✓	✓									
Imagen Editor [153]	Text, Mask	✓		✓			✓		✓		✓		
SmartMask [154]	Text	✓											
Uni-paint [155]	Text, Mask, Ref. Image	✓	✓	✓									
InstructPix2Pix [156]	Text	✓		✓	✓		✓	✓	✓				
MoEController [157]	Text	✓		✓	✓		✓	✓	✓				
FoI [158]	Text	✓		✓	✓		✓	✓					
LOFIE [159]	Text	✓		✓	✓							✓	
InstructDiffusion [160]	Text	✓	✓	✓			✓	✓	✓				
Emu Edit [161]	Text	✓	✓	✓	✓	✓			✓				
DialogPaint [162]	Text	✓	✓	✓	✓		✓						
Inst-Inpaint [163]	Text		✓										
HIVE [164]	Text	✓	✓	✓		✓	✓	✓					
ImageBrush [165]	Ref. Image, Seg. Map, Pose		✓	✓	✓			✓	✓				
InstructAny2Pix [166]	Ref. Image, Audio, Text	✓	✓		✓			✓					
MGIE [167]	Text		✓	✓									
SmartEdit [168]	Text		✓	✓									✓
iEdit [169]	Text	✓		✓		✓	✓	✓	✓			✓	
TDIELR [170]	Text	✓		✓					✓			✓	
ChatFace [171]	Text	✓				✓	✓						
UniTune [172]	Text	✓		✓	✓		✓	✓	✓				
Custom-Edit [173]	Text, Ref. Image			✓				✓					
KV-Inversion [174]	Text											✓	
Null-Text Inversion [175]	Text			✓	✓	✓	✓	✓	✓			✓	
DPL [176]	Text			✓									
DiffusionDisentanglement [177]	Text	✓		✓		✓	✓	✓	✓				
Prompt Tuning Inversion [178]	Text	✓		✓			✓	✓	✓			✓	

(下页继续)

TABLE 1 (continued)

Method	Condition(s)	Semantic Editing					Stylistic Editing			Structural Editing			
		Obj. Add.	Obj. Remo.	Obj. Repl.	Bg. Chg.	Emo. Expr. Mod.	Color Chg.	Style Chg.	Text. Chg.	Obj. Move.	Obj. Size. Chg.	Obj. Act. Chg.	Persp. /View. Chg.
StyleDiffusion [179] InST [180]	Text Text	✓		✓		✓		✓	✓				
DragonDiffusion [181] DragDiffusion [182] DDS [183] DiffuseIT [184] CDS [185] MagicRemover [186]	Dragging Points Dragging Points Text Text Text Text			✓		✓			✓	✓	✓	✓	
Imagic [187] LayerDiffusion [188] Forgedit [189] SINE [190]	Text Mask, Text Text Text	✓		✓			✓	✓		✓	✓	✓	✓
PRedItOR [191] ReDiffuser [192] Captioning and Injection [193] InstructEdit [194]	Text Text Text Text	✓		✓	✓		✓	✓	✓				
Direct Inversion [195] DDPM Inversion [196] SDE-Drag [197] LEDITS++ [198] FEC [199] EMILIE [200] Negative Inversion [201] ProxEdit [202] Null-Text Guidance [203] EDICT [204] AIDI [205] CycleDiffusion [206] InjectFusion [207] Fixed-point inversion [208] TIC [209] Diffusion Brush [210] Self-guidance [211]	Text Text Layout Text Text Text Text Text Ref. Image, Text Text Text Text Ref. Image Text Text, Mask Text, Mask Layout	✓	✓	✓	✓		✓	✓	✓	✓		✓	
P2P [212] Pix2Pix-Zero [213] MasaCtrl [85] PnP [214] TF-ICON [215] Object-Shape Variations [216] Conditional Score Guidance [217] EBMs [218] Shape-Guided Diffusion [219] HD-Painter [220]	Text Text Text, Pose, Sketch Text Text, Ref. Image, Mask Text Text Text Text, Mask Text, Mask	✓		✓	✓		✓	✓	✓			✓	
FISEdit [221] Blended Latent Diffusion [222] PFB-Diff [223] DiffEdit [224] RDM [225] MFL [226] Differential Diffusion [227] Watch Your Steps [228] Blended Diffusion [229] ZONE [230] Inpaint Anything [231]	Text Text, Mask Text, Mask Text Text Text Text, Mask Text Text, Mask Text, Mask Text, Mask	✓		✓	✓		✓	✓	✓				
The Stable Artist [232] SEGA [233] LEDITS [234] OIR-Diffusion [235]	Text Text Text Text	✓	✓	✓	✓		✓	✓	✓				

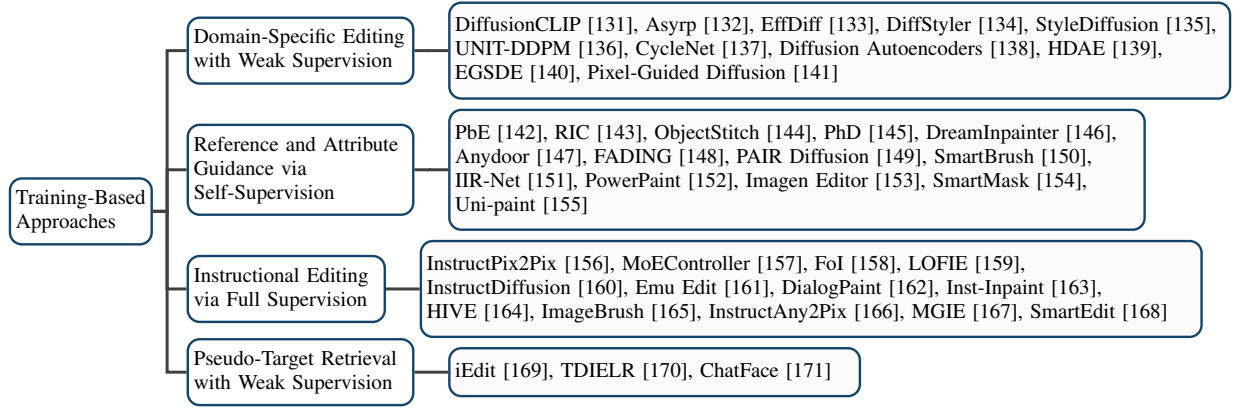


图 2：基于训练的图像编辑方法的分类。

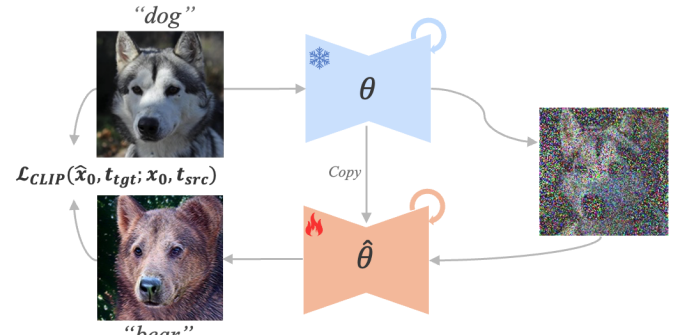
4 基于培训的方法

在基于扩散模型的图像编辑领域，基于训练的方法已获得显著的关注。这些方法不仅因其对扩散模型的稳定训练和对数据分布的有效建模而引人注目，而且因其在各种编辑任务中的可靠性能而引人注目。为了彻底研究这些方法，我们根据它们的应用范围、训练所需的条件和监督类型将它们分为四个主要组，如图 2 所示。此外，在每个主要组中，我们根据它们的核心编辑方法将方法分为不同的类型。这种分类说明了这些方法的范围，从有针对性的特定领域应用到更广泛的开放世界用途。

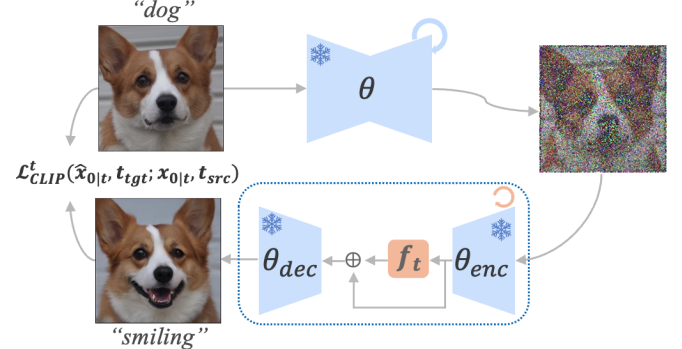
4.1 弱监督领域特定编辑

过去几年中，生成对抗网络 (GAN) 因其生成高质量图像的能力而在图像编辑领域得到了广泛探索。然而，扩散模型凭借其先进的图像生成能力，成为该领域的新焦点。扩散模型面临的一个挑战是在大型数据集上训练时需要大量的计算资源。为了解决这个问题，早期研究通过较小的专门数据集上的弱监督来训练这些模型。这些数据集高度集中在特定领域，如用于人脸处理的 CelebA [236] 和 FFHQ [2]、用于动物面部编辑和翻译的 AFHQ [237]、用于对象修改的 LSUN [238] 和用于风格转换的 WikiArt [239]。为了彻底理解这些方法，我们根据它们的弱监督类型对它们进行组织。

CLIP 指导。从基于 GAN 的方法 [240]、[241]（使用 CLIP [242] 指导带文本的图像编辑）中汲取灵感，一些研究将 CLIP 纳入扩散模型。一个关键的例子是 DiffusionCLIP [131]，它允许使用 CLIP 在训练域和新域中进行图像处理。具体来说，它首先使用 DDIM 反转将真实图像转换为潜在噪声，然后在反向扩散过程中微调预训练的扩散模型，以调整受源文本提示和目标文本提示之间的 CLIP 损失约束的图像属性。Asyrp [132] 不是对整个扩散模型进行微调，而是在内部关注一个语义潜在空间，称为 h -space，其中它定义了一个由小型神经网络参数化的附加隐式函数。然后，它在 CLIP 损失的指导下训练网络，同时保持扩散模型冻结。简化流程的视觉比较



(a) Simplified training pipeline of DiffusionCLIP [131]



(b) Simplified training pipeline of Asyrp [132]

图 3：两种代表性 CLIP 引导方法的训练流程比较：DiffusionCLIP [131] 和 Asyrp [132]。样本图像来自 AFHQ [237] 数据集上的 DiffusionCLIP [131] 和 Asyrp [132]。

DiffusionCLIP 和 Asyrp 如图 3 所示。为了解决 DiffusionCLIP 中多步优化耗时的问题，EffDiff [133] 引入了一种单步训练、高效处理的更快方法。

除了脸部编辑之外，上述方法主要关注的是 DiffStyler [134] 和 StyleDiffusion [135] 目标艺术风格迁移。DiffStyler 使用 CLIP 指令损失来对齐目标文本描述和生成的图像，并使用 CLIP 美学损失来增强视觉质量。另一方面，StyleDiffusion 引入了基于 CLIP 的风格解缠损失来改善风格与内容的协调。

循环正则化。由于扩散模型能够进行域转换，因此循环框架在

其中还探索了 CycleGAN [5] 等方法。例如, UNIT-DDPM [136] 在扩散模型中定义了一个双域马尔可夫链, 使用循环一致性来规范非配对图像到图像转换的训练。类似地, CycleNet [137] 采用 ControlNet [101] 和预训练的稳态扩散 [26] 作为文本调节的主干。它还在整个图像转换周期中采用一致性正则化, 其中包括从源域到目标域的正向转换和反向的后向转换。

投影和插值。GAN [243]、[244] 中经常使用的另一种技术是将两幅真实图像投影到 GAN 潜在空间中, 然后在它们之间进行插值, 以实现平滑的图像处理, 一些用于图像编辑的扩散模型也采用这种方法。例如, 扩散自编码器 [138] 引入了语义编码器, 将输入图像映射到语义上有意义的嵌入, 然后将其作为扩散模型重建的条件。在训练语义编码器和条件扩散模型后, 任何图像都可以投影到这个语义空间中进行插值。然而, HDAE [139] 指出, 这种方法往往会错过丰富的低级和中级特征。它通过增强框架来解决这个问题, 以分层利用语义编码器和基于扩散的解码器的由粗到细的特征, 旨在实现更全面的表示。

分类器引导。一些研究通过引入额外的预训练分类器进行引导, 提高了图像编辑性能。例如, EGSDE [140] 使用能量函数来指导现实的非配对图像到图像转换的采样。该函数由两个对数势函数组成, 分别由时间相关的域特定分类器和低通滤波器指定。对于细粒度图像编辑, 像素引导扩散 [141] 训练像素级分类器来估计分割图并用其梯度引导采样。

4.2 通过自我监督进行参考和属性指导

该类工作从单幅图像中提取属性或其他信息, 作为自监督训练基于扩散的图像编辑模型的条件, 可分为基于参考的图像合成和属性控制的图像编辑两类。

基于参考的图像合成。为了学习如何合成图像, PbE [142] 以自监督的方式进行训练, 使用图像中对象边界框内的内容作为参考图像, 使用边界框外的内容作为源图像。为了避免简单的复制粘贴解决方案, 它对参考图像应用了强增强, 基于边界框创建了任意形状的蒙版, 并使用 CLIP 图像编码器压缩参考图像的信息作为扩散模型的条件。

在此基础上, RIC [143] 将蒙版区域的草图作为训练的控制条件, 让用户通过草图精细调整参考图合成的效果。ObjectStitch [144] 设计了一个内容适配器, 以更好地保留参考图的关键身份信息。同时, PhD [145] 在冻结的预训练扩散模型上训练了一个修复和协调模块, 以有效地指导蒙版区域的修复。为了保留参考图的低级细节以供修复, DreamInpainter [146] 利用

U-Net 的下采样网络提取图像特征, 在训练过程中会给整幅图像添加噪声, 需要扩散模型学习如何在详细的文本描述的指导下恢复清晰的图像。此外, Anydoor [147] 使用视频帧中的图像对作为训练本来增强图像合成质量, 并引入了用于捕获身份特征、保留纹理和学习外观变化的模块。

属性控制的图像编辑。这类论文通常涉及用特定的图像特征作为控制条件来增强预训练的扩散模型, 以学习相应图像的生成。这种方法允许通过改变这些特定的控制条件来进行图像编辑。在对年龄-文本-面部对进行训练后, FADING [148] 通过空文本反转和注意力控制来编辑面部图像以进行年龄操纵。PAIR Diffusion [149] 将图像视为对象的集合, 学习调节每个对象的属性, 特别是结构和外观。SmartBrush [150] 使用不同粒度的蒙版作为控制条件, 使扩散模型能够根据文本和蒙版的形状修复蒙版区域。为了更好地保留与编辑文本无关的图像信息, IIR-Net [151] 在所需区域执行颜色和纹理擦除。然后将擦除后的图像用作扩散模型的控制条件之一。

4.3 全程监督下的教学编辑

使用指令(例如“摘下帽子”)来驱动图像编辑过程, 而不是使用已编辑图像的描述(例如“一只戴着帽子和微笑的小狗”)似乎更自然、更人性化, 也更能满足用户的需求。InstructPix2Pix [156] 是第一个学习按照人类指令编辑图像的研究。后续研究在模型架构、数据集质量、多模态性等方面对其进行了改进。因此, 我们首先描述 InstructPix2Pix, 然后根据其最突出的贡献对后续研究进行分类和展示。相应地, 这些基于指令的方法的通用框架如图 4 所示。

InstructPix2Pix 框架。使扩散模型能够根据指令编辑图像的主要挑战之一是构建指令-图像配对数据集。InstructPix2Pix 分两步生成这些图像对。首先, 给定一个图像标题(例如, “骑马的女孩的照片”), 它使用经过微调的 GPT-3 [245] 生成指令(例如, “让她骑龙。”)和编辑后的图像标题(例如, “骑龙的女孩的照片。”)。其次, 它采用稳态扩散和 Prompt-to-Prompt 算法 [212] 来生成编辑后的图像, 收集了超过 450,000 个训练图像对。然后, 它以完全监督的方式训练指令图像编辑扩散模型, 同时考虑输入图像和指令的条件。

模型架构增强。MoEController [157] 引入了混合专家 (MOE) 架构, 其中包括三位专门从事细粒度局部翻译、全局风格转换和复杂局部编辑任务的专家。另一方面, FoI [158] 利用 InstructPix2Pix 的隐式基础功能来识别和关注特定的编辑区域。它还采用跨条件注意力调制来确保每条指令都针对其相应的区域, 从而减少多条指令之间的干扰。

数据质量增强。LOFIE [159] 利用分割领域的最新进展, 提高了训练数据集的质量

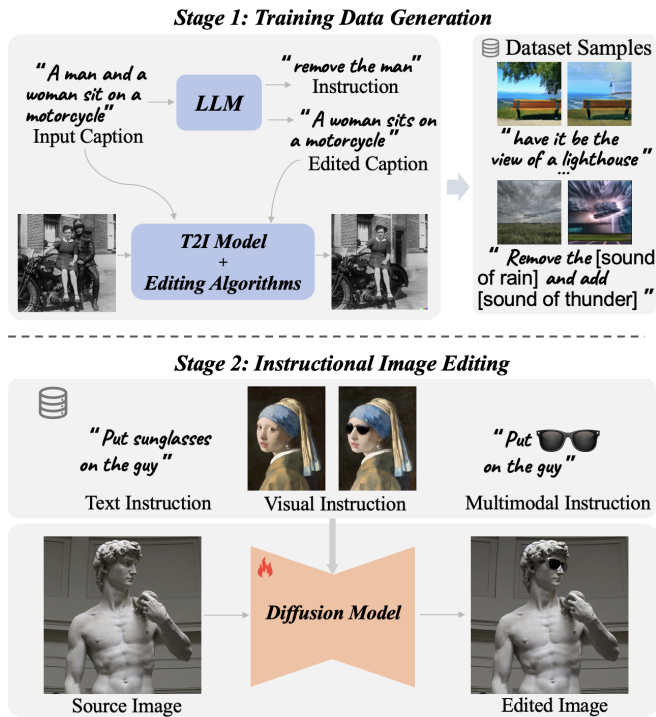


图 4：教学图像编辑方法的通用框架。样本图像来自 InstructPix2Pix [156]、InstructAny2Pix [166] 和 MagicBrush [249]。

[246]、思维链提示 [247] 和视觉问答 (VQA) [248]。MagicBrush [249] 从 Amazon Mechanical Turk (AMT) 聘请众包工人，使用 DALL-E 2 [25] 手动执行连续编辑。它包含 5,313 个编辑会话和 10,388 个编辑轮次，从而为教学图像编辑建立了全面的基准。InstructDiffusion [160] 是一个统一的框架，它将各种视觉任务视为人类直观的图像处理过程，即关键点检测、分割、图像增强和图像编辑。对于图像编辑，它不仅利用现有数据集，还使用对象移除和替换工具生成的额外数据以及通过从互联网上的真实 Photoshop 请求中收集图像编辑请求来扩充它们。Emu Edit [161] 也接受了图像编辑和识别数据的训练，利用包含 16 个不同任务的数据集，其中包含 1000 万个示例。该数据集使用 Llama 2 [250] 和上下文学习创建，以生成多样化且富有创意的编辑指令。在训练过程中，该模型结合其权重学习任务嵌入，仅使用少量示例即可高效适应新任务。

DialogPaint [162] 旨在提取多轮对话过程中用户的编辑意图并据此编辑图像。它采用 GPT-3 上的自指导技术 [251] 创建多轮对话数据集，并结合四个图像编辑模型生成指导性图像编辑数据集。此外，作者对 Blender 对话模型 [252] 进行了微调，以根据对话数据生成相应的编辑指令，然后驱动经过训练的指令编辑模型编辑图像。

Inst-Inpaint [163] 允许用户通过文本命令指定要从图像中删除的对象，而无需二进制掩码。它构建了一个名为 GQA-

Inpaint，基于图像和场景图数据集 GQA [253]。它首先从场景图中选择对象及其对应关系，然后使用 Detectron2 [254] 和 Detic [255] 提取这些对象的分割蒙版。之后，它使用 CRFill [256] 生成修复后的目标图像。编辑指令通过固定模板生成。构建数据集后，训练 Inst-Inpaint 执行指导性图像修复。

人工反馈增强学习。为了提高编辑图像与人工指导之间的一致性，HIVE [164] 在教学图像编辑中引入了人工反馈强化学习 (RLHF)。在获得 [156] 的基础模型后，在人工排名数据集上训练奖励模型。奖励模型的估计被整合到训练过程中，以微调扩散模型，使其与人工反馈保持一致。

视觉指令。ImageBrush [165] 被提出用于从一对说明所需操作的变换图像中学习视觉指令，并应用该指令来编辑新图像。该方法将示例图像、源图像和空白图像连接成网格，使用扩散模型根据示例图像提供的上下文信息迭代地对空白图像进行去噪。此外，还提出了一种视觉提示编码器，用于从视觉指令中提取特征以增强扩散过程。

利用多模态大规模模型。InstructAny2Pix [166] 使用户能够通过集成音频、图像和文本的指令来编辑图像。它使用多模态编码器 ImageBind [257] 将不同的输入转换为统一的潜在空间表示。Vicuna-7b [258] 是一个大型语言模型 (LLM)，它对多模态输入序列进行编码，以预测两个特殊标记作为将多模态输入与扩散模型的编辑结果对齐的条件。

MGIE [167] 将图像和指令以及多个 [IMG] 标记输入到多模态大型语言模型 (MLLM) LLaVA [259] 中。然后，它将倒数第二层的 [IMG] 隐藏状态投影到稳定扩散中 UNet 的交叉注意层。在训练过程中，LLaVA 和稳定扩散的权重会联合优化。

类似地，SmartEdit [168] 采用 LLaVA，并另外引入了双向交互模块 (BIM)，以便在复杂场景中执行图像编辑。它首先使用 QFormer [248] 将 MLLM 的隐藏状态与 CLIP 文本编码器对齐，然后通过 BIM 促进图像特征与 QFormer 输出之间的融合。此外，它利用与感知相关的数据（例如分割数据）来增强模型对空间和概念属性的理解。此外，它在复杂的理解和推理场景中结合了合成编辑数据，以激活 MLLM 的推理能力。

4.4 弱监督的伪目标检索

由于获取准确代表事实的编辑图像具有挑战性，此类方法旨在检索伪目标图像或直接使用 CLIP 分数 [242] 作为优化模型参数的目标。iEdit [169] 使用弱监督训练扩散模型，利用 CLIP 检索和编辑与编辑文本最相似的数据集图像，作为伪目标图像后编辑。此外，它将蒙版纳入图像编辑过程，以便通过 CLIPSeg [260] 进行局部保存。为了有效解决基于区域的

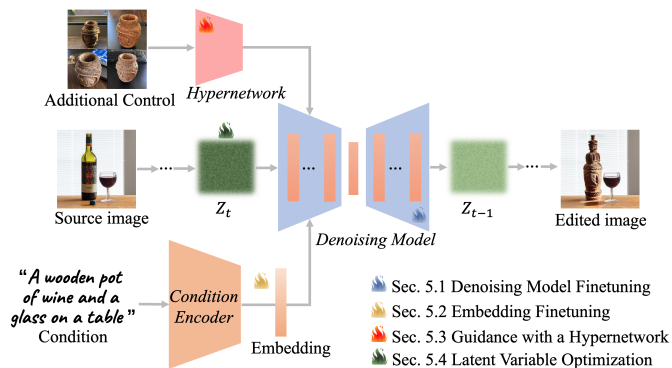


图 5：具有不同微调组件的测试时间微调框架。样本图像来自 Custom-Edit [173]。

图像编辑，TDIELR [170] 首先使用 DINO [261] 处理输入图像，以生成用于锚点初始化的注意图和特征。它学习区域生成网络 (RGN) 来选择最合适的区域提案。然后将选定的区域和文本描述输入到预训练的文本到图像模型中进行编辑。TDIELR 使用 CLIP 计算分数，评估文本描述和编辑结果之间的相似性，为 RGN 提供训练信号。此外，ChatFace [171] 也利用 CLIP 分数作为指标来学习如何在真实面部图像上进行编辑。

5 测试时间微调方法

在图像生成和编辑中，测试时微调代表着朝着精度和控制迈出的重要一步。本节探讨了增强图像编辑功能的各种微调策略（图 5）。如图 6 所示，这些方法包括微调整个去噪模型，以及专注于特定层或嵌入。我们研究了微调整个模型、针对特定参数和优化基于文本的嵌入的方法。此外，我们还讨论了超网络和直接图像表示优化的集成。这些方法共同展示了图像编辑中微调技术不断发展的复杂性和有效性，可满足广泛的编辑要求和用户意图。

5.1 去噪模型微调

去噪模型是图像生成和编辑的核心组件。直接对其进行微调是一种简单有效的方法。因此，许多编辑方法都基于此微调过程。其中一些涉及微调整个去噪模型，而另一些则侧重于微调模型中的特定层。

微调整个去噪模型。微调整个去噪模型可使模型更好地学习图像的特定特征并更准确地解释文本提示，从而使编辑更符合用户意图。UniTune [172] 在调整阶段对单个基础图像上的扩散模型进行微调，鼓励模型生成与基础图像相似的图像。在采样阶段，使用改进的采样过程来平衡对基础图像的保真度和与编辑提示的对齐。这包括从基础图像的噪声版本开始采样，并在采样过程中应用无分类器指导。Custom-Edit [173] 使用一小组参考图像来定制扩散模型，增强

编辑后的图像与参考图像的相似度，同时保持与源图像的相似度。去噪模型中的部分参数微调。一些方法专注于微调去噪模型的特定部分，例如自注意层、交叉注意层、编码器或解码器。这种类型的微调更精确，在很大程度上保留了预训练模型的功能并以此为基础。KV Inversion [174] 通过学习键 (K) 和值 (V)，设计了一种增强版的自注意，称为内容保留自注意 (CP-attn)。这有效地解决了在真实图像上进行动作编辑的问题，同时保留了原始图像的内容和结构。它为图像编辑提供了一种高效而灵活的解决方案，无需对模型进行微调或在大规模数据集上进行训练。

5.2 嵌入微调

许多微调方法选择以文本或空文本嵌入为目标进行细化，从而允许嵌入与生成过程更好地集成，以实现增强的编辑结果。

空文本嵌入微调。空文本嵌入微调的目的是解决 DDIM 反演 [18] 中重建失败的问题，从而提高与原始图像的一致性。在空文本反演 [175] 中，首先将 DDIM 反演应用于原始图像以获得反演轨迹。然后，在采样过程中，对空文本嵌入进行微调以减少采样轨迹和反演轨迹之间的距离，以便采样过程可以重建原始图像。这种方法的优点是 U-Net 权重和文本嵌入都不会改变，因此可以在不改变用户设置的目标提示的情况下提高重建性能。类似地，DPL [176] 使用泄漏注视损失动态更新文本提示中的名词，以解决交叉注意泄漏问题。对空文本嵌入进行微调以确保高质量的交叉注意图和准确的图像重建。

文本嵌入微调。对来自输入文本的嵌入进行微调可以增强图像编辑，使编辑后的图像更符合条件特征。DiffusionDisentanglement [177] 引入了一种简单的轻量级图像编辑算法，通过优化两个文本嵌入的混合权重来实现风格匹配和内容保存。这个过程涉及优化大约 50 个参数，优化后的权重在不同图像中具有很好的泛化能力。Prompt Tuning Inversion [178] 设计了一种基于文本的图像编辑的准确、快速的反演技术，包括重建阶段和编辑阶段。在重建阶段，它将输入图像的信息编码为可学习的文本嵌入。在编辑阶段，通过线性插值计算新的文本嵌入，将目标嵌入与优化后的嵌入相结合，以实现有效的编辑，同时保持高保真度。

5.3 超网络引导

除了传统的生成框架外，一些方法还结合了自定义网络，以更好地与特定的编辑意图保持一致。StyleDiffusion [179] 引入了一个映射网络，该网络将输入图像的特征映射到与文本提示的嵌入空间对齐的嵌入空间，从而有效地生成提示嵌入。交叉注意层是

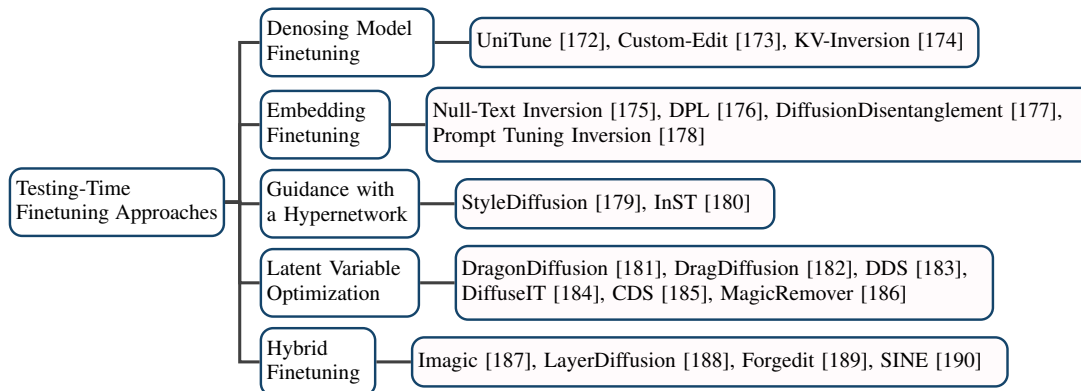


图 6：图像编辑的测试时间微调方法的分类。

用于将文本提示嵌入与图像特征表示相结合。这些层通过计算键、值和查询的注意力图来实现文本与图像的交互。InST [180] 在其文本反转部分集成了多层交叉注意力机制来处理图像嵌入。学习到的关键信息被转化为文本嵌入，可以将其看作是代表艺术品独特风格的“新词”，有效地表达了其独特的风格。

5.4 隐变量优化

直接优化图像的潜在变量也是微调过程中采用的一项技术。这种方法通过引入某些损失函数关系和某些中间层的特征来直接优化噪声潜在变量，而不是优化生成器的参数或嵌入的条件参数。利用预训练的扩散模型，大多数方法都获得了执行图像转换的能力，而无需配对训练数据。

人为引导的潜在变量优化。这种方法允许用户参与图像编辑过程，指导图像的生成。以 DragGAN [262] 为代表，这种交互式编辑过程使用户能够指定图像中的特定点并将其移动到新位置，同时保持图像的其余部分不变。DragGAN 通过优化 GAN 的潜在空间来实现这种编辑。随后，出现了基于扩散模型的发展，例如 DragonDiffusion [181]，它在扩散模型的中间特征中构建能量函数来指导编辑。这使得图像编辑指导可以直接通过图像特征进行，而不受文本描述的影响。DragDiffusion [182] 专注于在特定时间步骤而不是跨多个时间步骤优化扩散潜在表示。该设计基于以下观察：特定时间步骤的 U-Net 特征图提供了充足的语义和几何信息，以促进拖放编辑。

利用网络层和输入优化隐变量。一些优化方法利用从输入条件中得出的嵌入或网络特征来构建损失函数，从而能够直接优化隐变量。DDS [183] 方法利用两个图像-文本对：一个包含源图像及其描述性文本，另一个包含目标图像及其相应的描述性文本。DDS 计算这两个图像-文本对之间的差异，并通过这种比较得出损失。DiffuseIT [184] 中的损失函数还结合了 CLIP 模型的文本和图像编码器的使用

计算目标文本和源图像的相似度。CDS [185] 将对比较学习损失函数集成到 DDS 框架中，利用 LDM [26] 自注意层的空间丰富特征，通过对比损失计算来指导图像编辑。

5.5 混合微调

有些研究将上述各种微调方法结合起来，这些方法可以是连续的，即调整的各个阶段串联进行，也可以作为单个集成工作流程的一部分同时进行。这种复合微调方法可以实现有针对性且有效的图像编辑。

文本嵌入和去噪模型微调。Imagic [187] 分阶段实现其目标，首先将目标文本转换为文本嵌入，然后通过最小化嵌入和图像之间的差异来优化文本嵌入以重建输入图像。同时，对扩散模型进行微调以实现更好的图像重建。通过在优化的文本嵌入和目标文本之间进行线性插值找到中点，结合两者的特征。然后，微调后的扩散模型使用该嵌入生成最终的编辑图像。LayerDiffusion [188] 优化文本嵌入以匹配输入图像的背景，并采用分层扩散策略对模型进行微调，增强其保持主体和背景一致性的能力。Forgedit [189] 专注于快速图像重建和寻找适合编辑的文本嵌入，分别利用扩散模型的编码器和解码器来学习图像布局和纹理细节。

文本编码器和去噪模型微调。SINE [190] 最初对文本编码器和去噪模型进行微调，以更好地理解单个图像的内容和几何结构。它引入了一种基于块的微调策略，允许模型以任何分辨率生成图像，而不仅仅是预训练模型的固定分辨率。通过这种微调和块训练，SINE 能够处理单幅图像编辑任务，包括但不限于风格转换、内容添加和对象操作。

6 种无需训练和微调的方法

在图像编辑领域，无需训练和微调的方法以快速和低成本为前提，因为它们在整个编辑过程中不需要任何形式的训练（对数据集）或微调（对源图像）。本节将它们分为五类

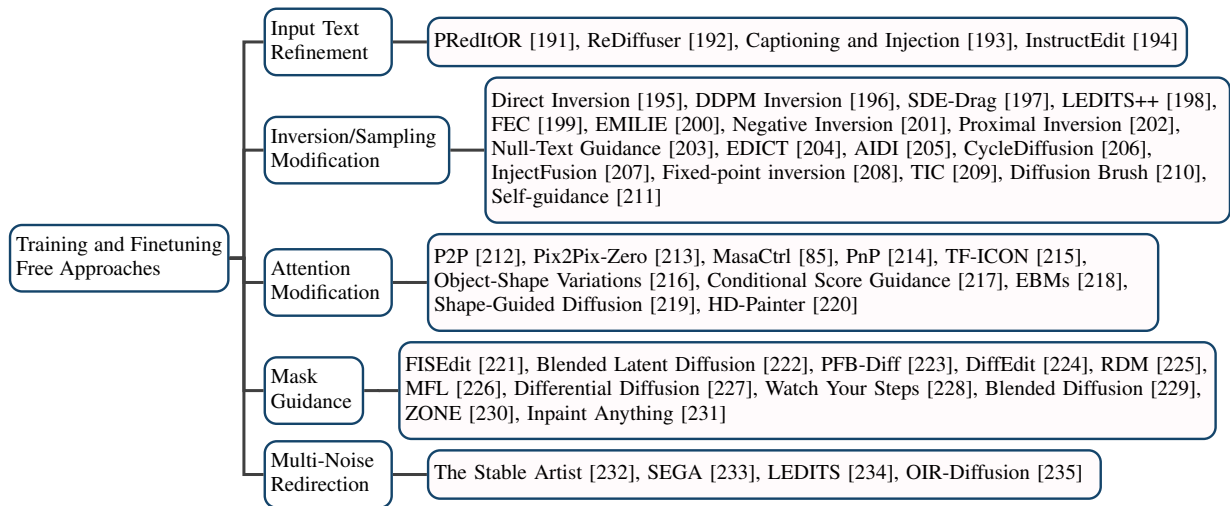


图 7：用于图像编辑的无需训练和微调的方法的分类。

如图 7 和图 8 所示，他们巧妙地利用了扩散模型中固有的原理来实现编辑目标。

6.1 输入文本细化

在图像编辑领域，输入文本细化标志着文本到图像转换机制的改进取得了重大进展。这种方法强调增强文本嵌入和简化用户输入，试图保证图像的修改既准确又符合给定文本的上下文。它支持概念修改和直观的用户指令，摆脱了复杂模型修改的必要性。

增强输入文本嵌入。这些方法专注于改进文本嵌入，以修改文本到图像的转换。PRedItOR [191] 通过利用扩散先验模型在 CLIP 图像嵌入空间中执行概念编辑来增强输入文本嵌入，从而实现更细致入微且更具情境感知的图像编辑。ReDiffuser [192] 通过再生学习生成丰富的提示，然后将其融合以指导编辑过程，确保在保持源图像结构的同时增强所需特征。此外，Captioning and Injection [193] 通过使用字幕模型（如 B LIP [263]）和提示注入技术 [193] 来增强输入文本嵌入，引入了一种用户友好的图像编辑方法。该方法允许用户提供最少的文本输入（例如单个名词），然后自动生成或优化详细提示，有效指导编辑过程，从而实现更准确、更符合情境的图像编辑。

基于指令的文本指导。此类别允许通过用户指令对图像编辑过程进行细粒度控制。它代表了向前迈出的重要一步，因为用户无需太多技术专业知识即可控制编辑过程。InstructEdit [194] 引入了一个框架，该框架由语言处理器、类似 SAM [246] 的分割模型和图像编辑模型组成。语言处理器将用户指令解释为分割提示和标题，随后利用这些提示和标题创建蒙版并指导图像编辑过程。这种方法展示了基于指令的文本指导在实现详细而精确的图像编辑方面的能力。

6.2 反演/采样修改

修改反演和采样公式是训练和微调免方法的常用技术 [195], [196], [201]。反演过程用于将真实图像反转到噪声潜在空间，然后采样过程用于在给定目标提示的情况下生成编辑结果。其中，DDIM 反演 [18] 是最常用的基线方法，尽管它经常导致重建源图像的失败 [196], [201], [204]。直接反演 [195] 开创了一种通过将源提示更改为目标提示来编辑真实图像的范式，展示了其处理各种任务的能力。然而，它仍然面临重建失败的问题。因此，一些方法 [196], [201], [204] 修改了反演和采样公式，以提高重建能力。

重建信息记忆。DDPM 反演 [196]、SDE-Drag [197]、LEDIT S++ [198]、FEC [199] 和 EMILIE [200] 都属于重建信息记忆类别；它们保存反演阶段的信息，并在采样阶段使用它来确保重建。DDPM 反演通过在反演阶段保留所有噪声潜变量来探索 DDPM 的潜在噪声空间，从而实现用户友好的图像编辑，同时确保重建性能。SDE-Drag 引入随机微分方程作为统一和改进的编辑框架，并展示其在拖动条件下编辑方面的优势。LEDITS++ 解决了 DDIM 反演中的低效率问题，提出了一种 DPM-Solver [89] 反演方法。FEC 专注于一致性增强，提供了三种在反演过程中保存信息的技术，以提高重建一致性。EMILIE 率先采用迭代多粒度编辑，提供迭代能力和多粒度控制，以实现所需的更改。这些方法通过记忆某些反演信息来确保重建性能，并在此基础上增强编辑操作的多样性。

在采样中使用空文本。负反转 [201]、Prox-Edit [202] 和空文本指导 [203] 都探讨了空文本（也称为负提示）信息对图像编辑的影响。负反转在采样过程中使用源提示作为负提示，实现与基于优化的方法相当的重建质量。ProxEdit 使用更大的无分类器指导尺度来解决 DDIM 反转重建中的退化问题，并提出了一种有效的

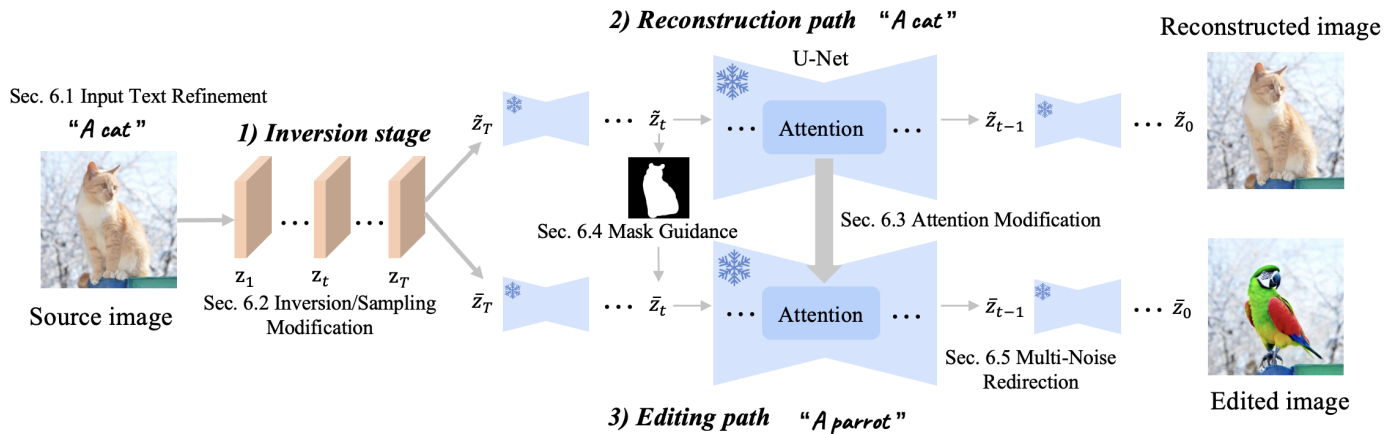


图 8：训练和微调自由方法的通用框架，其中指出了不同部分中描述的修改。示例图像来自 LEDITS++ [198]。

通过近端引导和相互自注意力控制来解决问题。同时，零文本引导 [203] 利用扰动方案将生成的图像转换为卡通，而无需训练。尽管这些方法对零文本的用途不同，但它们都从不同角度探索了零文本或负面提示设置在图像编辑中的作用。

单步多重噪声预测。EDICT[204]和AIDI[205]提出在单步采样过程中预测多重噪声，以解决重建失败的问题。EDICT从仿射耦合层[204]中汲取灵感，实现了真实图像的数学精确反演。它保持耦合噪声向量，能够在没有局部线性化假设的情况下实现精确的图像重建，其性能优于DDIM反演。另一方面，AIDI引入了加速迭代扩散反演，并通过在单步反演过程中预测多重噪声来提高重建精度。通过采用混合引导技术，AIDI在各种图像编辑任务中展示了有效的结果，无分类器引导等级较低，范围从1到3。

6.3 注意力机制修改

注意力修改方法增强了注意力层中的操作，这是最常见、最直接的无需训练的图像编辑方式。在稳定扩散的 U-Net 中，存在许多交叉注意力和自注意力层，这些层中的注意力和特征图包含大量语义信息。注意力修改的共同特点是需要辨别注意力层中的内在原理，然后通过修改注意力操作来利用它们。Prompt-to-Prompt (P2P) [212] 是注意力修改的先驱研究。

注意力图替换。P2P 通过将交叉注意力层确定为控制图像布局和提示词之间空间关系的关键，引入了一种直观的提示到提示编辑框架，该框架仅依赖于文本输入。如图 8 所示，给定重建和编辑路径，P2P 将编辑路径的注意力图替换为重建路径的相应注意力图，从而使编辑结果与源图像保持一致。与 P2P 不同，Pix2Pix-Zero [213] 消除了实际图像编辑中对用户定义文本提示的需求。它自主发现文本嵌入空间中的编辑方向，保留原始

内容结构。同时，它在 U-Net 中引入了交叉注意引导来保留输入图像的特征。注意特征替换。MasaCtrl [85] 和 PnP [214] 都强调替换注意特征以实现一致性。MasaCtrl 将自注意力转换为相互自注意力，它替换了自注意力层中的 Key 和 Value 特征以进行动作编辑。mask 引导的相互自注意力策略通过解决前景和背景的混淆进一步增强了一致性。PnP 通过空间特征操作和自注意力实现对生成结构的细粒度控制，直接注入指导图像特征。虽然两者都有助于特征替换，但 MasaCtrl 专注于自注意力层中的 Key 和 Value，而 PnP 强调自注意力层中的 Query 和 Key。

局部注意力图修改。TF-ICON [215] 框架和对象形状变化 [216] 共享局部注意力图修改的使用。TF-ICON 专注于跨域图像引导合成，将用户提供的对象无缝集成到视觉环境中，无需额外训练。它引入了对源图像自注意力的精确图像反转和变形映射的特殊提示。同时，对象形状变化 [216] 旨在生成描绘文本到图像工作流中特定对象形状变化的集合。它采用提示混合进行形状选择，并提出了一种使用自注意力层的定位技术。

注意分数指导。条件分数指导 [217] 和 EBM [218] 都致力于使用不同的注意分数指导。条件分数指导通过引入条件分数函数专注于图像到图像的翻译。该函数同时考虑源图像和文本提示，以选择性地编辑区域并保留其他区域。混合技术增强了未编辑区域和已编辑区域的融合，确保了高保真翻译。EBM 解决了与目标提示的语义不一致的问题。它引入了一个基于能量的自适应上下文控制框架，对上下文向量的后验进行建模以增强语义对齐。

6.4 口罩佩戴指导

基于扩散的图像编辑中的蒙版引导代表了一种增强图像编辑的技术。它包括通过选择性处理来提高去噪效率的技术、针对性图像编辑的精确蒙版自动生成技术，以及

掩模引导区域聚焦，以确保局部修改与特定感兴趣区域准确对齐。这些方法利用掩模来引导和改进采样过程，从而提供更高的精度、速度和灵活性。

掩模增强去噪效率。这些方法利用掩模来提高基于扩散的图像编辑的效率。[221] – [223] 方法优先提高速度，同时确保高质量的结果。通过掩模选择性地处理图像区域，它们有效地减少了计算需求并提高了整体效率。FISEdit [221] 引入了一种支持缓存的稀疏扩散模型，该模型利用输入文本上的微小修改和输出图像上受影响区域之间的语义映射。它自动识别受影响的图像区域，并利用缓存的未改变区域的特征图来加速推理过程。混合潜在扩散 [222] 利用文本到图像的潜在扩散模型 [26]，该模型通过在低维潜在空间内运行来加速扩散过程，并有效地消除了扩散模型的每个时间步长上对 CLIP [242] 梯度计算的要求。PFB-Diff [223] 通过多级特征混合将文本条件生成的内容无缝集成到目标图像中，并在交叉注意力层引入注意掩蔽机制，将特定单词的影响限制在所需区域，从而提高背景编辑的性能。

蒙版自动生成。这些方法通过自动生成蒙版实现了精确的图像编辑。DiffEdit [224] 通过自动生成隔离相关区域以进行修改的蒙版来简化语义编辑。这种选择性编辑方法可以保护未编辑区域并保持其语义完整性。RDM [225] 引入了一种区域感知扩散模型，该模型无缝集成了蒙版，可根据文本驱动的指导自动精确定位和编辑感兴趣的区域。MFL [226] 提出了一种两阶段无蒙版训练范式，专门用于文本引导的图像编辑。在第一阶段，根据源提示获得统一的蒙版，然后基于扩散模型，使用提供的蒙版和目标提示生成多个候选图像。

蒙版引导的区域聚焦。这一系列研究使用蒙版作为导航工具，将采样过程引导至特定的感兴趣区域。这些方法 [227]、[228] 在定位编辑方面表现出色，确保修改与文本提示或编辑要求中指定的指定区域精确对齐。差分扩散 [227] 开创了一种区域感知编辑范式，其中使用蒙版来定制源图像每个区域内变化的强度。这允许进行细粒度的控制，从而促进精致的细节和艺术表现。“小心你的脚步” [228] 利用蒙版作为相关性图来精确定位所需的编辑区域，确保只编辑最相关的区域，同时保留原始图像的其他区域。

6.5 多噪声重定向

多噪声重定向是一次性预测多个不同方向的噪声，然后将它们重定向为单个噪声的过程。这种重定向的优点是它能够使用单个噪声同时统一多个不同的编辑方向，从而更有效地满足用户的编辑需求。

语义噪声控制。这些方法在语义上指导采样过程中的噪声重定向，并增强细粒度

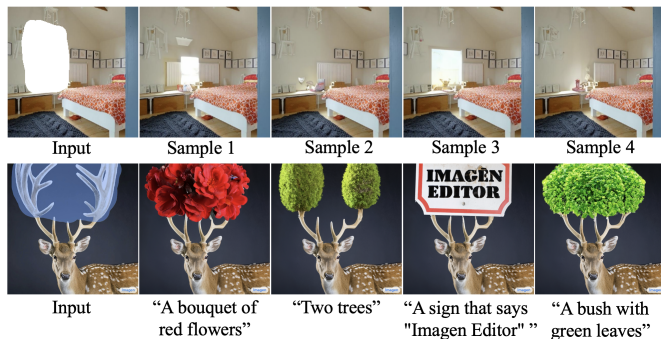


图 9：传统上下文驱动修复（上）与多模态条件修复（下）的视觉比较。两行样本分别来自 Palette [58] 和 Imagen Editor [153]。

控制图像内容。Stable Artist [232] 和 SEGA [233] 展示了使用语义指导来引导扩散过程的能力。它们通过控制噪声重定向过程中的各种语义方向来促进图像合成的精细编辑。LEDITS [234] 通过将 DDPM 反转 [196] 与 SEGA 相结合来应用语义噪声控制，并促进了多种编辑方向，包括对象编辑和整体风格变化。

对象感知噪声重定向。OIR-Diffusion [235] 引入了对象感知反转和重组方法，通过确定每个样本的最佳反转步骤并将已编辑区域与其他区域无缝集成，可以对图像中的特定对象进行细粒度编辑。它显著保持了未编辑区域的保真度，并在编辑各种对象属性方面取得了显著的效果，尤其是在复杂的多对象场景中。

7 绘画与外绘

图像修复和图像修复通常被视为图像编辑的子任务，它们占据着独特的地位，具有不同的目标和挑战。为了更好地解释，我们将它们分为两种主要类型（见图 9），详见第 7.1 节和第 7.2 节。

7.1 传统的上下文驱动修复

7.1.1 Supervised Training

基于监督训练的修复和修复方法通常使用成对的损坏图像和完整图像从头开始训练扩散模型。Palette [58] 基于条件扩散模型开发了一个统一的图像到图像转换框架，并在四个图像到图像转换任务上分别训练该框架，即着色、修复、修复和 JPEG 恢复。它利用低质量参考图和 $t-1$ 步骤的去噪结果的直接连接作为 t 步骤噪声预测的条件。SUD² [264] 建立并扩展了 SUD 框架 [265]，集成了其他策略，如相关性最小化、噪声注入，特别是使用去噪扩散模型来解决半监督学习中配对训练数据的稀缺问题。

7.1.2 Zero-Shot Learning

上下文优先集成。该方法旨在从图像的未损坏部分中提取固有结构和纹理，作为

在每个步骤中，它们互补地生成被遮蔽的损坏区域中的内容，从而保持全局内容一致性。具有代表性的是，Repaint [266] 以预训练的无条件扩散模型为基础，通过在每个反向步骤中从扩散模型中采样被遮蔽的区域并从给定图像中采样未遮蔽的区域来修改标准去噪策略。GradPaint [267] 也利用预训练的扩散模型来估计被遮蔽的图像区域的内容。然后，它使用自定义损失函数来测量与未遮蔽部分的一致性，并根据损失的梯度更新生成的内容以确保上下文一致性。

退化分解。图像修复可以看作是一般线性逆问题的一种特殊应用，其表示为：

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \quad (13)$$

其中 $\mathbf{H}$ 为退化算子（例如，由 0 和 1 组成的掩码矩阵，表示修复过程中图像的损坏区域）， $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ 。目标是从损坏的图像 $\mathbf{y}$ 中恢复原始图像 $\mathbf{x}$ 。一些工作通过对 $\mathbf{H}$ 进行分解来实现它。DDRM [66] 利用预训练的去噪扩散模型和 $\mathbf{H}$ 的奇异值分解 (SVD)，将这种恢复转换为 $\mathbf{H}$ 谱空间内高效的迭代扩散过程。不同的是，DDNM [68] 通过应用范围零空间分解来解决图像逆问题，其中图像通过退化矩阵 $\mathbf{H}^\dagger$ 的伪逆分为两个不同的分量。

后验估计。为了解决公式 13 中提出的一般噪声线性逆问题，一些研究 [70]、[268] – [272] 侧重于估计后验分布 $p(\mathbf{x}|\mathbf{y})$ ，利用基于贝叶斯定理的无条件扩散模型。它在逆扩散过程的每一步计算条件后验 $p(\mathbf{x}_t|\mathbf{y})$ 。利用贝叶斯定理，可以推导出： $p(\mathbf{x}_t|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{y})}$ 。然后计算对数似然的相应梯度：

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \quad (14)$$

其中，得分函数 $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ 可以使用预训练得分网络 $\mathbf{s}_\theta(\mathbf{x}_t, t)$ 来估计。然后，对 $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ 的估计简化为对未知项 $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$ 的估计。两个具有代表性的早期研究 MCG [70] 和 DPS [268] 用 $p_t(\mathbf{y}|\hat{\mathbf{x}}_0)$ 近似后验 $p_t(\mathbf{y}|\mathbf{x}_t)$ ，其中 $\hat{\mathbf{x}}_0$ 是 $\mathbf{x}_t$ 的无噪声期望，使用 Tweedie 公式将其形式化为 $\hat{\mathbf{x}}_0 = E[\hat{\mathbf{x}}_0|\mathbf{x}_t]$ 。因此，等式 14 变为：

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) &= \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \\ &\approx -\rho \nabla_{\mathbf{x}_t} \|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}_0\|_2^2 + \mathbf{s}_\theta(\mathbf{x}_t, t), \end{aligned} \quad (15)$$

其中 $\rho = \frac{1}{\sigma^2}$ 是步长。MCG 进一步通过公式 15 在更新后将投影添加到测量子空间，作为偏离理想数据一致性的纠正措施。PIGDM [269] 通过引入退化函数 $\mathbf{H}$ 的 Moore-Penrose 伪逆 $\mathbf{H}^\dagger$ ，扩展了公式 14 中的后验估计框架，以解决线性、非线性和可微逆问题。GDP [270] 给出了不同的近似值，公式如下：

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t) \approx -s \nabla_{\mathbf{x}_t} \mathcal{L}(\mathbf{H}\hat{\mathbf{x}}_0 + \mathbf{n}, \mathbf{y}) - \lambda \nabla_{\mathbf{x}_t} \mathcal{Q}(\hat{\mathbf{x}}_0), \quad (16)$$

其中， $\mathcal{L}$ 、 $\mathcal{Q}$ 、 s 和 λ 分别是图像距离度量、质量损失、控制引导幅度的比例因子和调整图像质量的比例因子。出于类似的目标，CoPaint [271] 使用最终生成图像的一步估计来确定性地预测 $\hat{\mathbf{x}}_0$ ，以使在每个中间去噪步骤中修复误差的计算变得易于处理。

7.2 多模态条件修复

7.2.1 Random Mask Training

传统的图像修复侧重于根据周围环境填充图像中的缺失区域，通常缺乏对内容的精确控制。随着文本到图像扩散模型的进步，通过注入用户提供的额外多模态条件（例如文本描述、分割图和参考图作为指导），可以克服这一限制。这些模型（例如 Stable Diffusion 和 Imagen）也可以适用于修复任务，其中在反向扩散过程中将噪声背景替换为原始图像的噪声版本。然而，由于采样期间全局上下文可见性有限，这种修改有时会导致不令人满意的样本。为了解决这个问题，GLIDE [34] 和 Stable Inpainting（来自 Stable Diffusion [26] 的修复专家 v1.5）等模型进一步微调了专门用于修复的预训练文本到图像模型。他们使用随机生成的蒙版以及蒙版图像和完整图像标题，使模型能够在训练期间利用未蒙版区域的信息。作为开创性的工作，它们为后续研究提供了启发性的见解。

7.2.2 Precise Control Conditioning

随机掩码的一个基本问题是，它可能会覆盖与文本提示无关的区域，导致模型忽略提示，尤其是当掩码区域较小或仅覆盖物体的一部分时。为了对修复的内容提供更精确的控制，已经提出了几种方法。SmartBrush [150] 引入了一个精度因子，通过对精确的实例掩码应用高斯模糊来生成各种类型的掩码，从精细到粗糙。它允许用户在粗蒙版（包括蒙版内的所需物体）和详细蒙版（紧密勾勒出物体形状）之间进行选择。

Imagen Editor [153] 通过微调扩展了 Imagen [27]，采用级联扩散模型进行文本引导修复。它使用三个卷积下采样图像编码器将图像和掩码上下文集成到每个扩散阶段。这里的掩码是由物体检测器 SSD Mobilenet v2 [273] 即时生成的，而不是随机生成的。

PbE [142] 和 PhD [145] 专注于使用参考图像进行用户自定义修复。具体来说，PbE 会丢弃文本，而仅使用 CLIP [242] 从参考图像中提取主题信息。参考图是从修复图像本身裁剪出来的，并进行增强以支持自监督训练。相比之下，PhD 与 PbE 一样，在训练流程中集成了文本和参考图像。

Uni-paint [155] 首先对输入图像中未遮罩区域上的空文本进行稳定扩散微调，然后使用各种引导模式（包括无条件、文本驱动、笔画驱动、样例驱动修复以及它们的组合）通过不同的条件集成技术在推理过程中执行多模态修复。对于内部修复，

PVA [274] 通过 Transformer 模块对一些与蒙版图像具有相同身份的参考图像进行编码，从而保留身份。输出图像嵌入与 CLIP 文本嵌入相结合，然后通过并行视觉注意机制作为稳定修复模型的额外条件。SmartMask [154] 利用语义非模态分割数据为无蒙版对象插入创建高质量注释。它训练一个扩散模型，从这些注释中预测实例图，包括语义图和场景上下文。在推理过程中，它使用全景分割来创建语义图，从而指导对象插入蒙版的预测。

7.2.3 Pretrained Diffusion Prior Exploiting

尽管基于训练的方法取得了重大进展，但仍然存在挑战，特别是在收集大规模真实数据和模型训练所需的大量资源方面。因此，一些研究将重点转向特定的修复任务，通过整合各种技术来利用预训练的文本到图像扩散模型的生成能力。

混合扩散 [229] 使用 CLIP 计算预训练无条件扩散模型的去噪图像嵌入与引导文本嵌入之间的差异，以优化屏蔽区域内的生成。然后，在每个采样步骤中，它将噪声屏蔽区域与输入图像未屏蔽部分的相应噪声版本进行空间混合。混合潜伏扩散 [222] 通过将这项技术应用于 LDM [26] 中，进一步提高了效率。

Inpaint Anything [231] 将 SAM [246] 和现有的修复模型（如 LaMa [275] 和 Stable Inpainting [222]）结合为一个整体框架，允许用户通过简单的点击移除、替换或填充图像的任何部分，在图像编辑中提供了高度的用户友好性和灵活性。MagicRemover [186] 从空文本反转 [175] 开始，将图像投影到嘈杂的潜在代码中，然后在采样过程中优化后续代码。它引入了基于注意的引导策略和分类器优化算法，以更少的步骤提高修复过程的精度和稳定性。

HD-Painter [220] 专为高分辨率图像修复而设计，分为两个阶段：图像修复和修复专用超分辨率。第一阶段在稳定扩散的潜在空间中进行，其中提示感知的内向注意力机制取代了原来的自我注意力，并采用了额外的重新加权注意力分数指导。此修改可确保生成的内容与输入提示更好地对齐。第二阶段使用稳定扩散超分辨率变体，对低分辨率修复图像进行调节并混合蒙版。

7.3 修饰

图像外绘与修复类似但不完全相同，其目的是生成无缝扩展图像边界的新像素。现有的 T2I 模型（如稳定扩散 [26] 和 DALL-E [25]）可以推广到解决此任务，因为它们在包括各种大小和形状的图像的大量数据集上进行训练。它通常被认为是具有类似实现的一种特殊形式的修复。例如，Palette [58] 通过将裁剪图像与其原始版本以监督方式直接组合作为输入来训练扩散模型。PowerPaint [152] 将修复和相关任务分为四个特定类别，将外绘视为一种情境感知

图像修复。它通过可学习的任务特定提示和微调策略增强了稳定扩散，从而精确地将模型引导至不同的目标。

8 基准与评估

8.1 基准测试构建

在前面的部分中，我们深入探讨了基于扩散模型的图像编辑方法的方法论方面。除了这些分析之外，评估这些方法、检查它们在各种编辑任务中的能力也至关重要。然而，现有的图像编辑基准测试有限，不能完全满足我们调查中确定的需求。例如，EditBench [153] 主要针对文本和蒙版引导的修复，而忽略了涉及全局编辑（如风格迁移）的更广泛任务。Ted Bench [187] 虽然扩大了任务范围，但缺乏详细的说明，而这对于评估依赖文本说明而非描述的方法至关重要。此外，EditVal 基准测试 [276] 虽然试图提供更全面的任务和方法覆盖，但它受到 MS-COCO 数据集 [277] 中图像质量的限制，这些图像通常分辨率较低且模糊。

为了解决这些问题，我们引入了 EditEval，这是一个旨在评估基于一般扩散模型的图像编辑方法的基准。EditEval 包含一个精心策划的数据集，其中包含 50 张高质量图像，每张图像都附有文本提示。EditEval 评估从表 1 中选择的 7 个常见编辑任务的性能。此外，我们提出了 LMM 分数，这是一种定量评估指标，它利用大型多模态模型 (LMM) 的功能来评估不同任务的编辑性能。除了 LMM 分数提供的客观评估外，我们还进行了用户研究以纳入主观评估。EditEval 的构建及其应用的细节如下所述。

8.1.1 Task Selection

在选择评估任务时，我们考虑了表 1 所示的能力。我们发现，大多数方法能够处理语义和风格任务，但在结构编辑方面却遇到了挑战。潜在的原因是，大多数编辑方法所依赖的许多当前 T2I 扩散模型难以实现准确的空间感知 [278]。例如，它们通常会生成不一致的图像，提示中包含“左边”、“下面”或“后面”等词语。考虑到这些因素和实际应用，我们选择了七个常见任务作为基准：添加对象、替换对象、移除对象、替换背景、更改风格、更改纹理和更改动作，旨在对从简单的对象编辑到复杂的场景变化的编辑方法性能进行全面评估。

8.1.2 Dataset Construction

对于图片选择，我们从 Un-splash 的专业照片在线库中手动选择了 50 张图片，确保主题和场景的多样性。这些图片被裁剪成正方形格式，与大多数编辑模型的输入比例一致。然后，我们将选定的图片分为 7 组，每组都与上述特定编辑任务之一相关。

在为每个图像生成提示时，我们使用 LMM 创建描述图像内容的源提示、概述编辑预期结果的目标提示，以及

相应的指令来指导编辑过程。通过为最广泛使用的 LMM 之一 GPT-4V 提供一个包含示例对和几条指令的详细模板，可以简化此步骤。示例对由任务指示器（例如“移除对象”）、源图像、源描述（例如“未整理的床上的一杯咖啡”）、目标描述（例如“未整理的床”）和指令（例如“移走那杯咖啡”）组成。除了这个例子，我们还向 GPT-4V 提供了明确的指令，说明我们对生成源描述、目标描述以及我们上传的新图像和任务指示器的指令的期望。这种方法确保 GPT-4V 生成的提示和指令不仅与指定的编辑任务一致，而且还保持场景和主题多样性。在 GPT-4V 生成初始提示和指令后，我们会进行细致的检查，以确保每个提示和指令集都是具体的、清晰的，并且直接适用于相应的图像和任务。

最终数据集（包括所选图像、其源和目标提示以及编辑说明）可在随附的 GitHub 存储库中找到。

8.1.3 Metric Design and Selection

图像编辑领域传统上依赖 CLIP-Score [279] 作为主要的定量评估指标。尽管 CLIPScore 在评估图像与相应文本提示之间的对齐方面很有效，但它在具有许多细节和特定空间关系的复杂场景中可能会遇到困难 [276]。这一限制促使我们需要一种更通用、更全面的指标，以应用于更广泛的图像编辑任务。因此，我们提出了 LMM Score，这是一种新的指标，它利用了大型多模态模型 (LMM) 的高级视觉语言理解能力。

为了制定该指标，我们首先指示 GPT-4 构思一个定量指标并概述一个框架，该框架允许通过适当的用户提示对一般图像编辑任务进行客观评估。根据 GPT-4 的建议，评估框架包含以下元素：使用源图像 I_{src} 及其文本描述 t_{src} 、由 N 方法生成的编辑图像集合 $I_{tgt}^i, i = 1, 2, \dots, N$ 、编辑提示 t_{edit} 、编辑指令 t_{inst} 和任务指示器 o 。该标准整合了四个关键因素：

- *Editing Accuracy*：评估编辑后的图像 I_{tgt}^i 与指定的编辑提示 t_{edit} 和指令 t_{inst} 的遵循程度，衡量编辑的精度
- *Contextual Preservation*：评估编辑后的图像 I_{tgt}^i 对不需要进行任何更改的源图像 I_{src} 的上下文的保留程度
- *Visual Quality*：评估编辑后的图像 I_{tgt}^i 的整体质量，包括分辨率、无伪影、色彩准确度、清晰度等因素。
- *Logical Realism*：评估编辑后的图像 I_{tgt}^i 在遵循自然法则（例如光照一致性、纹理连续性等）方面的逻辑真实感。

至于评估，每张编辑后的图像 I_{tgt}^i 都会根据这些因素进行评估，得出四个子分数（范围从 1 到 10），分别表示为 S_{acc} 、 S_{pre} 、 S_{qua} 和 S_{real} 。总分 S_{LMM} 是这些子分数的加权平均值，以反映整体编辑质量：

$$S_{LMM} = 0.4S_{acc} + 0.3S_{pre} + 0.2S_{qua} + 0.1S_{real}. \quad (17)$$

GPT-4 提出的这种加权公式旨在平衡各个评估因素的相对重要性。

按照所提出的框架，我们使用 LMM（本研究中具体为 GPT-4V）来进行评估。GPT-4V 遵循用户预定义的说明来计算每个编辑图像的四个评估因素的子分数，从而获得 S_{LMM} 的值（如公式 17 所示）。为了使 LMM 分数的应用更加方便用户，我们提供了一个包含预定义说明和必要材料的模板，该模板可在我们的 GitHub 存储库中访问，从而方便其在图像编辑评估中的采用。

除了 LMM Score 提供的客观评价外，我们还进行了用户研究以收集主观反馈。这项研究吸引了 50 名来自不同背景的参与者，以确保观点的广泛性。每位参与者都会看到源图像、其描述、编辑提示和说明以及一系列编辑后的图像。然后，他们被要求根据 LMM Score 使用的相同四个评估因素对每张编辑后的图像进行评分。每个因素的分数从 1 到 10。然后汇总参与者的分数以计算每张图像的总用户分数 (S_{User})，计算方法与公式 17 中相同。这项用户研究的结果补充了 LMM Score 评估，为不同编辑方法的性能提供了全面的评估。

8.2 评估

8.2.1 Method Selection

为了评估每项编辑任务，我们从表 1 中精心选择了 4 到 8 种方法，涵盖了一系列方法，包括基于训练的方法、测试时微调的方法以及无需训练和微调的方法。为了确保公平和一致的比较，我们的选择标准如下：该方法必须只需要文本条件、具有处理特定任务的能力，并且有可供实施的开源代码。我们排除了特定领域的方法，以避免它们在不同领域中的适用性有限所带来的限制。这些选定的方法将在后续小节中详细说明。

8.2.2 Comparative Analysis

性能比较。为了对 7 个选定的编辑任务中选定的方法进行全方面评估，我们计算了所有评估样本的 S_{LMM} 和 S_{User} 的平均值和标准差，如图 10 所示。颜色条表示从最小到最大的分数范围。从结果中，我们可以得到几个见解。首先，没有一种方法可以在所有任务中胜过其他方法。例如，虽然 LEDITS++ [198] 在物体移除方面表现出了值得称赞的性能（图 10 (c)），但在风格变化方面表现不佳（图 10 (e)）。其次，大多数方法在每个任务中都表现出很大的分数波动范围，最小值和最大值之间的跨度很大， S_{LMM} 和 S_{User} 的标准差也很大。这种多变性表明这些方法不稳健，表明方法的性能可能依赖于样本，并且可能无法始终如一地满足各种场景的编辑要求。第三，有些方法在某些任务上确实表现出稳定且令人印象深刻的性能。例如，空文本反转 (Null-Text) [175] 在对象替换、背景变化和纹理变化中始终表现出较高的得分分布，表明其可靠性。

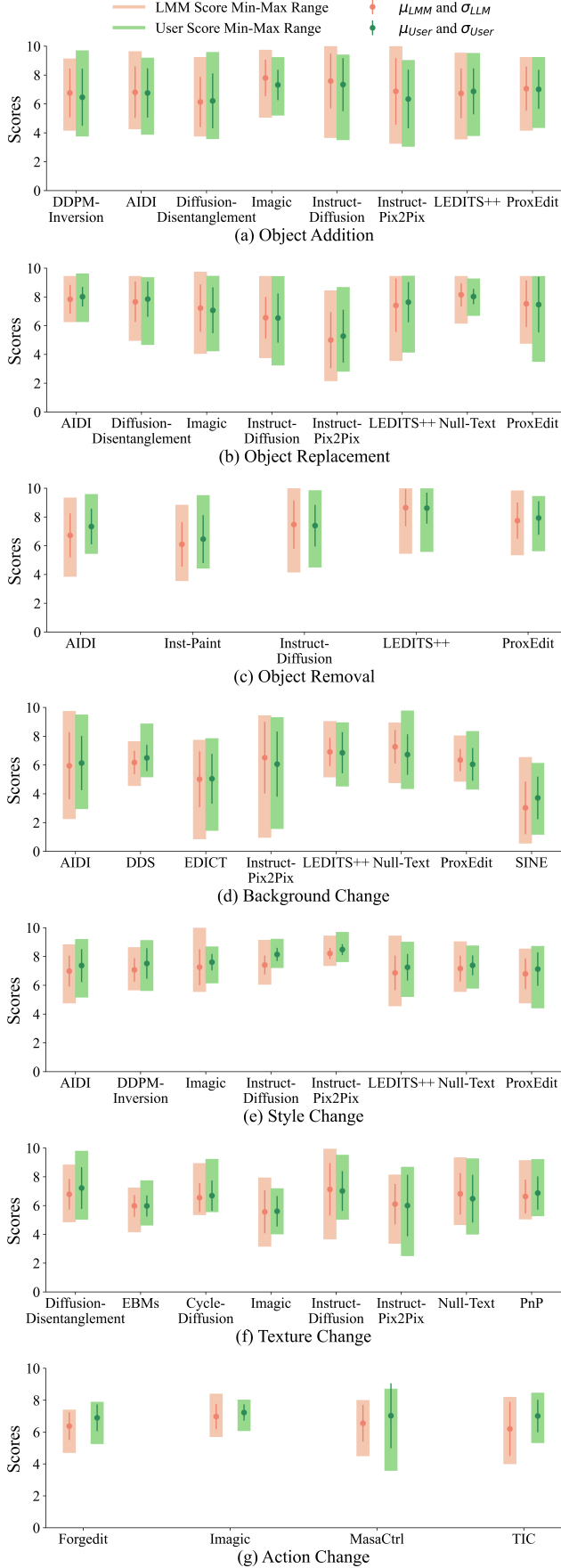


图 10：各种方法在 7 个选定的编辑任务中的定量表现。 μ 和 σ 分别表示分数的平均值和标准差。



图 11：LMM 分数和用户研究之间的皮尔逊相关系数。

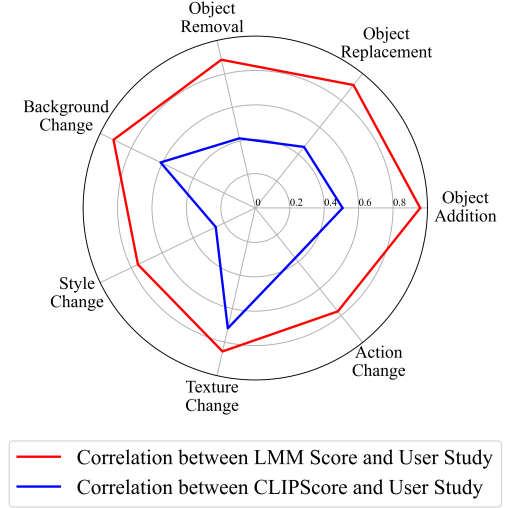


图 12：LMM Score/CLIPScore 与用户研究的皮尔逊相关系数比较。

编辑质量对于一些实际应用来说非常重要。最后, S_{LMM} 和 S_{User} 的平均值对于每个任务中的每个方法都非常接近, 这表明 LMM 分数很好地反映了用户的偏好。尽管方法的性能会有所波动, 但 LMM 分数与任务间用户感知之间的一致性表明 LMM 分数是图像编辑的可靠指标。除了定量结果外, 我们还在图 13 中提供了几个用于定性比较的视觉示例。

LMM 分数与用户研究之间的相关性。为了检验 LMM 分数的有效性, 我们研究了它与用户研究之间的相关性。具体来说, 我们首先计算收集到的 4 个因素的用户数据的 4 个子分数的平均值。然后, 在每个编辑任务中, 我们计算每个样本的 LMM 子分数和用户研究子分数之间每个因素的皮尔逊相关系数。最后, 对这些系数在所有样本中取平均值, 以获得每个任务每个因素的总体相关性。如图 11 所示, 结果显示在编辑任务之间存在显著的一致性, 表明客观的 LMM 分数评估与主观的人为判断之间存在很强的一致性。

LMM 分数与 CLIPScore。为了进一步评估 LMM 分数的有效性, 我们计算了常用指标 CLIPScore [279] 与用户分数之间的皮尔逊相关系数, 以供比较。CLIPScore 是使用指定的

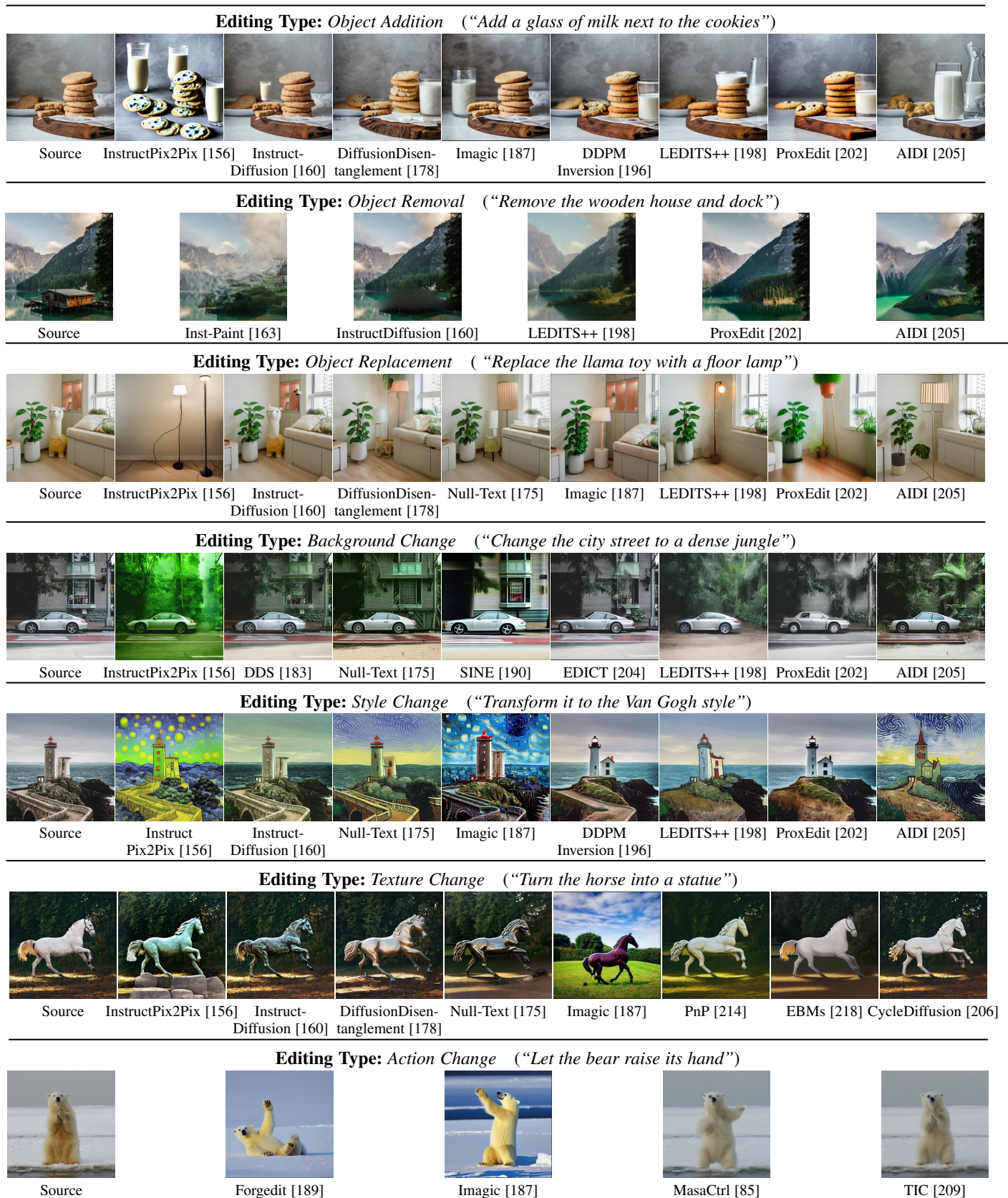


图 13：7 种选定编辑类型的视觉比较。

文本提示 t_{edit} , 针对源样本 I_{src} 和在特定编辑任务中通过方法 i 生成的相应编辑图像 I_{tgt}^i [279]。在 7 个选定任务中获得所有样本的所有 CLIPScore 值后, 我们将计算这些分数与所有样本的总体用户分数之间的皮尔逊相关系数。与 LMM 分数不同, CLIPScore 不考虑特定因素, 因此直接与 S_{User} 进行比较。然后对这些系数取平均值, 以得出每个任务的总体相关系数。同时, 我们计算所有样本的总体 LMM 分数 S_{LMM} 和总体用户分数之间的皮尔逊相关系数, 并对每个任务取平均值。所有 S_{LMM} 分数都归一化到 $[0, 1]$ 以匹配 CLIPScore 的范围。图 12 将这些比较可视化, 并表明对于所有这些编辑任务, LMM Score 都比 CLIPScore 与用户判断保持了更密切的相关性, 表明它更准确地反映了用户评价以及 LMM Score 在评估各种图像编辑任务和方法方面的有效性。

9 挑战和未来方向

尽管使用扩散模型进行图像编辑取得了成功, 但仍存在一些需要在未来工作中解决的限制。

少步模型推理。大多数基于扩散的模型在推理过程中需要大量步骤才能获得最终图像, 这既耗时又耗计算, 给模型部署和用户体验带来了挑战。为了提高推理效率, 人们研究了少步或一步生成扩散模型 [90]、[280]、[281]。该领域的最新方法通过从预训练的强扩散模型中提取知识来减少步骤数, 从而使少步模型可以模仿强模型的行为。一个更有趣但更具挑战性的方向是直接开发少步模型而不依赖于预训练模型, 例如一致性模型 [282]、[283]。

高效模型。训练一个能够产生真实结果的扩散模型需要大量计算, 并且需要大量高质量的数据。这种复杂性使得开发用于图像编辑的扩散模型非常具有挑战性。为了降低训练成本, 最近的研究设计了更高效的网络架构作为扩散模型的主干 [284]、[285]。此外, 另一个重要方向是只训练一部分参数或冻结原始参数并在预训练的扩散模型上添加一些新的层 [113]、[286]。

复杂对象结构编辑。现有作品可以在编辑图像时合成逼真的颜色、样式或纹理。然而, 在处理手指、徽标和场景文本等复杂结构时, 它们仍然会产生明显的伪影。人们已经尝试解决这些问题。以前的方法通常使用负面提示, 例如“六个手指, 腿不好等”, 以使模型避免生成此类图像, 这在某些情况下是有效的, 但不够稳健 [287]。最近的作品开始使用布局、边缘或密集标签作为编辑图像全局或局部结构的指导 [101]、[103]。

复杂的光照和阴影编辑。编辑物体的光照或阴影仍然具有挑战性, 这需要准确估计场景中的光照条件。以前的研究, 如 Total Relighting [288], 使用网络组合来估计前景物体的法线、反照率和阴影, 以获得逼真的重新照明效果。最近, 基于扩散的

提出了一些模型来编辑面部光照 (DiFaReli [289])。然而, 利用预训练扩散模型的强光先验来编辑肖像或一般物体仍然是一个开放领域 [290]。同样, ShadowDiffusion [62] 探索了基于扩散的阴影合成, 可以生成视觉上令人愉悦的物体阴影。然而, 使用扩散模型在不同背景条件下准确编辑物体的阴影仍然是一个尚未解决的问题。

图像编辑的不稳健性。现有的基于扩散的图像编辑模型可以在给定的某些条件下合成逼真的视觉内容。然而, 它们在许多现实场景中仍然失败了 [187]。这个问题的根本原因是模型不能准确地对条件分布空间中的所有可能样本进行建模。如何改进模型以始终如一地生成无伪影的内容仍然是一个挑战。有几种方法可以缓解这个问题。首先, 扩大模型训练的数据以覆盖具有挑战性的场景。这是一种有效但成本高昂的方法。在某些情况下, 收集足够数量的数据甚至非常具有挑战性, 例如医学图像、视觉检查数据等。其次, 调整模型以接受更多条件, 如结构指导 [101]、3D 感知指导 [291] 和文本指导, 以实现更可控和确定的内容创建。第三, 采用迭代细化或多阶段训练, 逐步改进模型的初始结果 [57]、[266]、[292]。

忠实评估指标。准确的评估对于图像编辑至关重要, 以确保编辑的内容与给定的条件很好地一致。然而, 尽管一些定量指标如 FID [293]、KID [294]、LPIPS [295]、CLIP 分数 [242]、PSNR 和 SSIM 已被用于此任务作为参考, 但大多数现有工作仍然严重依赖用户研究来为视觉结果提供相对准确的感知评估, 这既不高效也不可扩展。忠实的定量评估指标仍然是一个悬而未决的问题。最近, 已经提出了更准确的量化物体感知相似性的指标。DreamSim [296] 考虑布局、姿势和语义内容来测量两幅图像的中级相似度, 其表现优于 LPIPS。同样, 前景特征平均 (FFA) [297] 提供了一种简单而有效的方法来测量物体的相似性, 而不管其姿势、视点、光照条件或背景如何。在本文中, 我们还借助 LMM 提出了一种有效的图像编辑指标 LMM Score。

10 结论

我们对基于扩散模型的图像编辑方法进行了广泛的概述, 从多个角度研究了该领域。我们的分析首先根据学习策略将 100 多种方法分为三大类: 基于训练的方法、测试时微调的方法以及无需训练和微调的方法。然后, 我们将图像编辑任务分为三个不同的类别: 语义编辑、风格编辑和结构编辑, 总共包含 12 种特定类型。我们探索了这些方法及其对提高编辑性能贡献。在我们的图像编辑基准 EditEval 中对 7 项任务以及最近的最先进方法进行了评估。此外, 还引入了一个新的指标 LMM 分数来对这些方法进行比较分析。在总结我们的评论时, 我们强调了图像编辑领域的广阔潜力并提出了未来研究的方向。

参考

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville 和 Y. Bengio, “生成对抗网络”, 载于 *NeurIPS*, 第 27 卷, 2014 年。[2] T. Karras, S. Laine 和 T. Aila, “基于风格的生成对抗网络生成器架构”, 载于 *CVPR*, 第 4401-4410 页, 2019 年。[3] M. Mirza 和 S. Osindero, “条件生成对抗网络”, *arXiv preprint arXiv:1411.1784*, 2014 年。[4] P. Isola, J.-Y. Zhu, T. Zhou, 和 A. A. Efros, “Image-to-image Translation with Conditional Adversarial Networks”, in *CVPR*, pp. 1125–1134, 2017 年。[5] J.-Y. Zhu, T. Park, P. Isola 和 A. A. Efros, “使用循环一致对抗网络进行非配对图像到图像转换”, 载于 *ICCV*, 第 2223-2232 页, 2017 年。[6] A. Brock, J. Donahue 和 K. Simonyan, “用于高保真自然图像合成的大规模 GAN 训练”, 载于 *ICLR*, 2019 年。[7] D. P. Kingma 和 M. Welling, “自动编码变分贝叶斯”, *arXiv preprint arXiv:1312.6114*, 2013 年。[8] A. Van Den Oord, N. Kalchbrenner 和 K. Kavukcuoglu, “像素递归神经网络”, 载于 *ICML*, 第 1747-1756 页, 2016 年。[9] D. Rezende 和 S. Mohamed, “带正则化的变分推理流”, 载于 *ICML*, 第 1530-1538 页, 2015 年。[10] A. Van Den Oord, O. Vinyals, “神经离散表征学习”, 载于 *NeurIPS*, 第 30 卷, 2017 年。[11] G. Papamakarios, T. Pavlakou 和 I. Murray, “用于密度估计的掩蔽自回归流”, 载于 *NeurIPS*, 第 30 卷, 2017 年。[12] P. Esser, R. Rombach 和 B. Ommer, “Taming transformers for high-resolution image synthesis”, 载于 *CVPR*, 第 12873-12883 页, 2021 年。[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen 和 I. Sutskever, “Zero-shot text-to-image generation”, 载于 *ICML*, 第 8821-8831 页, 2021 年。[14] J. Zhao, M. Mathieu 和 Y. LeCun, “基于能量的生成对抗网络”, *arXiv preprint arXiv:1609.03126*, 2016 年。[15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan 和 S. Ganguli, “使用非平衡热力学的深度无监督学习”, 载于 *ICML*, 2015 年。[16] J. Ho, A. Jain 和 P. Abbeel, “去噪扩散概率模型”, 载于 *NeurIPS*, 第 3 卷, 第 6840-6851 页, 2020 年。[17] A. Q. Nichol 和 P. Dhariwal, “改进的去噪扩散概率模型”, 载于 *ICML*, 第 8162-8171 页, 2021 年。[18] J. Song, C. Meng 和 S. Ermon, “去噪扩散隐式模型”, 载于 *ICLR*, 2021 年。[19] A. Hyvärinen 和 P. Dayan, “通过分数匹配估计非归一化统计模型”, 载于 *JMLR*, 第 6 卷, 第 4 期, 2005 年。[20] P. Vincent, “分数匹配与去噪自动编码器之间的联系”, *Neural computation*, 第 23 卷, 第 7 期, 第 1661-1674 页, 2011 年。[21] Y. Song 和 S. Ermon, “通过估计数据分布梯度进行生成建模”, 载于 *NeurIPS*, 第 32 卷, 2019 年。[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon 和 B. Poole, “通过随机微分方程进行基于分数的生成建模”, 载于 *ICLR*, 2021 年。[23] Y. Song 和 S. Ermon, “基于分数的生成模型的改进训练技术”, 载于 *NeurIPS*, 第 33 卷, 第 12438-12448 页, 2020 年。[24] J. Shi, C. Wu, J. Liang, X. Liu 和 N. Duan, “Divae: 使用去噪扩散解码器进行真实感图像合成”, *arXiv preprint arXiv:2206.00386*, 2022 年。[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu 和 M. Chen, “使用剪辑潜变量的分层文本条件图像生成”, *arXiv preprint arXiv:2204.06125*, 2022 年。[26] 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”, *CVPR*, 2022 年。[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipoor, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., “具有深度语言理解的真实感文本到图像扩散模型”, *NeurIPS*, 2022 年。[28] J. Ho 和 T. Salimans, “无分类器扩散指导”, *arXiv preprint arXiv:2207.12598*, 2022 年。[29] G. Batzolis, J. Stanczuk, C.-B. Schönlieb 和 C. Etmann, “基于分数的扩散模型的条件图像生成”, *arXiv preprint arXiv:2111.13606*, 2021 年。[30] F. Bao, C. Li, J. Zhu 和 B. Zhang, “Analytic-dpm: 扩散概率模型中最优逆方差的分析估计”, *ICLR*, 2022 年。[31] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi 和 T. Salimans, “用于高保真图像生成的级联扩散模型”, *JMLR*, 第 23 卷, 第 1 期, 第 2249-2281 页, 2022 年。[32] P. Dhariwal 和 A. Nichol, “扩散模型在图像合成方面击败了 GAN”, *NeurIPS*, 第 34 卷, 第 8780-8794 页, 2021 年。[33] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach 和 T. Darrell, “免费获得更多控制! 具有语义扩散引导的图像合成”, 载于 *WACV*, 第 289–299 页, 2023 年。[34] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgraw, I. Sutskever 和 M. Chen, “Glide: 使用文本引导的扩散模型实现照片级真实感图像生成和编辑”, 载于 *ICML*, 第 16784–16804 页, 2022 年。[35] R. Rombach, A. Blattmann 和 B. Ommer, “使用检索增强扩散模型进行文本引导的艺术图像合成”, *arXiv preprint arXiv:2207.13038*, 2022 年。[36] A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping 和 T. Goldstein, “冷扩散: 无噪声反转任意图像变换”, *arXiv preprint arXiv:2208.09392*, 2022 年。[37] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho 和 T. Salimans, “论引导扩散模型的提炼”, *CVPR*, 第 14297-14306 页, 2023 年。[38] H. Phung, Q. Dao 和 A. Tran, “小波扩散模型是快速且可扩展的图像生成器”, *CVPR*, 第 10199-10208 页, 2023 年。[39] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi 和 D. J. Fleet, “视频传播模型”, 载于 *NeurIPS*, 2022 年。[40] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gefni, et al., “制作视频: 无需文本视频数据的文本转视频生成”, 载于 *ICLR*, 2023 年。[41] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al., “Imagen 视频: 使用传播模型生成高清视频”, *arXiv preprint arXiv:2210.02303*, 2022 年。[42] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. 雅各布斯, J.-B. 黄, M.-Y. Liu 和 Y. Balaji, “保留您自己的相关性: 视频扩散模型的噪声先验”, 载于 *ICCV*, 2023 年。[43] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler 和 K. Kreis, “对齐您的潜在: 使用潜在扩散模型进行高分辨率视频合成”, 载于 *CVPR*, 2023 年。[44] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu 和 J. Feng, “Magicvideo: 使用潜在扩散模型高效生成视频”, *arXiv preprint arXiv:2211.11018*, 2022 年。[45] S. Yin, C. Wu, H. Yang, J. Wang, X. Wang, M. Ni, Z. Yang, L. Li, S. Liu, F. Yang, et al., “Nuwa-xl: 扩散优于扩散, 用于生成超长视频”, *arXiv preprint arXiv:2303.12346*, 2023 年。[46] P. Esser, J. Chiu, P. Atighehchian, J. Granskog 和 A. Germanidis, “具有扩散模型的结构和内容引导的视频合成”, *ICCV*, 2023 年。[47] J. An, S. Zhang, H. Yang, S. Gupta, J.-B. Huang, J. Luo 和 X. Yin, “Latent-shift: 具有时间移位的潜在扩散, 可实现高效的文本到视频生成”, *arXiv preprint arXiv:2304.08477*, 2023 年。[48] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang 和 S. Zhang, “Modelscope 文本到视频技术报告”, *arXiv preprint arXiv:2308.06571*, 2023 年。[49] X. Li, W. Chu, Y. Wu, W. Yuan, F. Liu, Q. Zhang, F. Li, H. Feng, E. Ding 和 J. Wang, “Videogen: 一种用于高清文本到视频生成的参考引导潜在扩散方法”, *arXiv preprint arXiv:2309.00398*, 2023 年。[50] Y. He, T. Yang, Y. Zhang, Y. Shan 和 Q. Chen, “用于任意长度的高保真视频生成的潜在视频扩散模型”, *arXiv preprint arXiv:2211.13221*, 2022 年。[51] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, et al., “Lavie: 使用级联潜在扩散模型生成高质量视频”, *arXiv preprint arXiv:2309.15103*, 2023 年。[52] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, 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的一次性调整”, *ICCV*, 2023 年。[53] J. Lv, Y. Huang, M. Yan, J. Huang, J. Liu, Y. Liu, Y. Wen, X. Chen 和 S. Chen, “Gpt4motion: 通过面向 blender 的脚本编写文本到视频生成中的物理动作 gpt 规划”, *arXiv preprint arXiv:2311.12631*, 2023 年。[54] D. J. Zhang, J. Z. Wu, J.-W. Liu, R. Zhao, L. Ran, Y. Gu, D. Gao 和 M. Z. Shou, “Show-1: 结合像素和潜在扩散模型进行文本到视频生成”, *arXiv preprint arXiv:2309.15818*, 2023 年。[55] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilián, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al., “稳定的视频”

- 扩散：将潜在视频扩散模型扩展到大型数据集”，*arXiv preprint arXiv:2311.15127*, 2023 年。[56] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh 和 I. Misra, “Emu 视频：通过显式图像条件分解文本到视频生成”，*arXiv preprint arXiv:2311.10709*, 2023 年。[57] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet 和 M. Norouzi, “通过迭代细化实现图像超分辨率”，*IEEE TPAMI*, 第 45 卷, 第 4, 第 4713-4726 页, 2022 年。[58] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet 和 M. Norouzi, “调色板：图像到图像的扩散模型”，载于 *ACM SIGGRAPH*, 第 1-10 页, 2022 年。[59] O. Özdenizci 和 R. Legenstein, “使用基于块的去噪扩散模型在恶劣天气条件下恢复视力”，*IEEE TPAMI*, 2023 年。[60] S. Shang, Z. Shan, G. Liu 和 J. Zhang, “Resdiff：结合 cnn 和扩散模型实现图像超分辨率”，*arXiv preprint arXiv:2303.08714*, 2023 年。[61] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen 和 B. Zhang, “连续超分辨率的隐式扩散模型”，载于 *CVPR*, 第 10021-10030 页, 2023 年。[62] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister 和 B. Wen, “阴影扩散：当退化先验遇到扩散模型进行阴影去除时”，载于 *CVPR*, 第 14049-14058 页, 2023 年。[63] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund 和 T. B. Schön, “使用均值回归随机微分方程进行图像恢复”，载于 *ICML*, 2023 年。[64] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang 和 L. Van Gool, “Diffir：用于图像恢复的有效扩散模型”，*arXiv preprint arXiv:2303.09472*, 2023 年。[65] J. Choi, S. Kim, Y. Jeong, Y. Gwon 和 S. Yoon, “Ilvr：去噪扩散概率模型的条件方法”，载于 *CVPR*, 第 14367-14376 页, 2021 年。[66] B. Kavar, M. Elad, S. Ermon 和 J. Song, “去噪扩散恢复模型”，载于 *NeurIPS*, 卷. 35, 第 23593-23606 页, 2022 年。[67] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lyu, C. Chen 和 S. Chen, “Wavedm：基于小波的图像恢复扩散模型”，*IEEE TMM*, 2024 年。[68] Y. Wang, J. Yu 和 J. Zhang, “使用去噪扩散零空间模型的零样本图像恢复”，*ICLR*, 2023 年。[69] Z. Yue 和 C. C. Loy, “Difface：具有扩散误差收缩的盲人脸恢复”，*arXiv preprint arXiv:2212.06512*, 2022 年。[70] H. Chung, B. Sim, D. Ryu 和 J. C. Ye, “使用流形约束改进逆问题的扩散模型”，*NeurIPS*, 第 35, 第 25683-25696 页, 2022 年。[71] H. Chung, J. Kim, M. T. McCann 和 M. L. Klasky 和 J. C. Ye, “扩散后验采样用于一般噪声逆问题”，*ICLR*, 2023 年。[72] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu 和 D. Merhof, “用于医学图像分析的扩散模型：综合调查”，*arXiv preprint arXiv:2211.07804*, 2022 年。[73] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu 和 Y.-G. Jiang, “视频扩散模型综述”，*arXiv preprint arXiv:2310.10647*, 2023 年。[74] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang 和 Z. Chen, “图像恢复和增强的扩散模型——综合综述”，*arXiv preprint arXiv:2308.09388*, 2023 年。[75] B. B. Moser, A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio 和 A. Dengel, “扩散模型、图像超分辨率及一切：综述”，*arXiv preprint arXiv:2401.00736*, 2024 年。[76] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui 和 M.-H. Yang, “扩散模型：方法和应用的综合综述”，*ACM Computing Surveys*, 第 56 卷, 第 4, 第 1-39 页, 2023 年。[77] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng 和 S. Z. Li, “生成扩散模型综述”，*arXiv preprint arXiv:2209.02646*, 2022 年。[78] F.-A. Croitoru, V. Hondru, R. T. Ionescu 和 M. Shah, “视觉中的扩散模型：一项调查”，*IEEE TPAMI*, 2023 年。[79] A. Ulhaq, N. Akhtar 和 G. Pogrebnia, “视觉的有效扩散模型：一项调查”，*arXiv preprint arXiv:2210.09292*, 2022 年。[80] C. Zhang, C. Zhang, M. Zhang 和 I. S. Kweon, “生成人工智能中的文本到图像扩散模型：一项调查”，*arXiv preprint arXiv:2303.07909*, 2023 年。[81] T. Zhang, Z. Wang, J. Huang, M. M. Tasnim 和 W. Shi, “基于扩散的图像生成模型调查：问题及其解决方案”，*arXiv preprint arXiv:2308.13142*, 2023 年。[82] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. H. Bernardino, E. R. Chan, T. Dekel, A. Holynski, A. Kanazawa, *et al.*, “视觉计算扩散模型的最新研究成果”，*arXiv preprint arXiv:2310.07204*, 2023 年。[83] H. Koo 和 T. E. Kim, “结构化数据生成扩散模型的综合调查”，*ArXiv, abs/2306.04139 v2*, 2023 年。[84] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu 和 S. Ermon, “Sdedit：使用随机微分方程进行图像合成和编辑”，载于 *ICLR*, 2022 年。[85] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie 和 Y. Zheng, “Masactrl：无需调整的相互自注意力控制，实现一致的图像合成和编辑”，*arXiv preprint arXiv:2304.08465*, 2023 年。[86] J. Ho 和 T. Salimans, “无分类器扩散指导”，载于 *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021 年。[87] C.-H. Chao, W.-F. Sun, B.-W. Cheng, Y.-C. Lo, C.-C. Chang, Y.-L. Liu, Y.-L. Chang, C.-P. Chen 和 C.-Y. Lee, “去噪似然得分匹配用于基于条件得分的数据生成”，载于 *ICLR*, 2022 年。[88] T. Karras, M. Aittala, T. Aila 和 S. Laine, “阐明基于扩散的生成模型的设计空间”，载于 *NeurIPS*, 第 35 卷, 第 26565-26577 页, 2022 年。[89] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li 和 J. Zhu, “Dpm-solver：一种用于扩散概率模型采样的快速 ode 求解器，大约 10 个步骤”，载于 *NeurIPS*, 第 35, 第 5775-5787 页, 2022 年。[90] T. Salimans 和 J. Ho, “扩散模型快速采样的渐进式蒸馏”，载于 *ICLR*, 2022 年。[91] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan 和 B. Guo, “文本到图像合成的矢量量化扩散模型”，载于 *CVPR*, 第 10696-10706 页, 2022 年。[92] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna 和 R. Rombach, “Sdxl：改进用于高分辨率图像合成的潜在扩散模型”，*arXiv preprint arXiv:2307.01952*, 2023 年。[93] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, *et al.*, “Pixart- α ：用于照片级真实感文本到图像合成的扩散变压器的快速训练”，*arXiv preprint arXiv:2310.00426*, 2023 年。[94] X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhennde, X. Wang, A. Dubey, *et al.*, “Emu：使用大海捞针增强图像生成模型”，*arXiv preprint arXiv:2309.15807*, 2023 年。[95] K. Sohn, N. Ruiz, K. Lee, D. C. Chin, I. Blok, H. Chang, J. Barber, L. Jiang, G. Entis, Y. Li, *et al.*, “Stylerdrop：任意风格的文本到图像生成”，*arXiv preprint arXiv:2306.00983*, 2023 年。[96] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, *et al.*, “ediffi：文本到图像的差异生成”模型与一组专家去噪器”，*arXiv preprint arXiv:2211.01324*, 2022 年。[97] S. Ge, T. Park, J.-Y. Zhu 和 J.-B. Huang, “使用富文本进行富有表现力的文本到图像生成”，载于 *CVPR*, 第 7545-7556 页, 2023 年。[98] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li 和 Y. J. Lee, “Gligen：基于开放集的文本到图像生成”，载于 *CVPR*, 第 22511-22521 页, 2023 年。[99] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh 和 Y. Taigman, “Make-a-scene：基于人类先验的基于场景的文本到图像生成”，载于 *ECCV*, 第 89-106 页, 2022 年。[100] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried 和 X. Yin, “Spatext：可控图像生成的空间文本表示”，载于 *CVPR*, 第 18370-18380 页, 2023 年。[101] L. Zhang, A. Rao 和 M. Agrawala, “为文本到图像扩散模型添加条件控制”，载于 *ICCV*, 第 3836-3847 页, 2023 年。[102] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan 和 K.-Y. K. Wong, “Uni-controlnet：文本到图像扩散模型的一体化控制”，*arXiv preprint arXiv:2305.16322*, 2023 年。[103] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, *et al.*, “Unicontrol：一种用于在野外控制视觉生成的统一扩散模型”，*arXiv preprint arXiv:2305.11147*, 2023 年。[104] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao 和 J. Zhou, “Composer：具有可组合条件的创意和可控图像合成”，*ICML*, 2023 年。[105] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan 和 X. Qie, “T2i-adapter：学习适配器以挖掘更多文本到图像扩散模型的可控能力”，*arXiv preprint arXiv:2302.08453*, 2023 年。

- [106] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik and D. Cohen-or, “一图胜千言：使用文本反转个性化文本到图像生成”, 载于 *ICLR*, 2023 年. [107] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman, “Dreambooth: 微调文本到图像扩散模型以实现主题驱动生成”, 载于 *CVPR*, 第 22500-22510 页, 2023 年. [108] N. Kumari, B. Zhang, R. Zhang, E. Shechtman and J.-Y. Zhu, “文本到图像传播的多概念定制”, 载于 *CVPR*, 第 1931-1941 页, 第 2023 页. [109] A. Voronov, M. Khoroshikh, A. Babenko and M. Ryabinin, “这种损失有参考价值吗? 通过跟踪目标动态实现更快的文本到图像定制”, 载于 *NeurIPS*, 2023 年. [110] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang and W. Zuo, “Elite: 将视觉概念编码到文本嵌入中以实现定制的文本到图像生成”, *arXiv preprint arXiv:2302.13848*, 2023 年. [111] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou and Y. Cao, “锥体: 扩散模型中的概念神经元用于定制生成”, 载于 *ICML*, 2023 年. [112] W. Chen, H. Hu, Y. Li, N. Rui, X. Jia, M.-W. Chang and W. W. Cohen, “通过学徒学习实现主题驱动的文本到图像生成”, *arXiv preprint arXiv:2304.00186*, 2023 年. [113] J. Shi, W. Xiong, Z. Lin and H. J. Jung, “Instantbooth: 无需测试时微调的个性化文本到图像生成”, *arXiv preprint arXiv:2304.03411*, 2023 年. [114] Y. Tewel, R. Gal, G. Chechik and Y. Atzmon, “用于文本到图像个性化的密钥锁定一级编辑”, 在 *ACM SIGGRAPH*, 第 1-11 页, 2023 年. [115] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou and W. Zhu, “Disen-booth: 用于主题驱动的文本到图像生成的身份保留解缠结调整”, *arXiv preprint arXiv:2305.03374*, 卷. 2, 2023 年. [116] J. Gu, Y. Wang, N. Zhao, T.-J. Fu, W. Xiong, Q. Liu, Z. Zhang, H. Zhang, J. Zhang, H. Jung, et al., “Photoswap: 图像中的个性化主题交换”, *arXiv preprint arXiv:2305.18286*, 2023 年. [117] Z. Yuan, M. Cao, X. Wang, Z. Qi, C. Yuan and Y. Shan, “Customnet: 文本到图像扩散模型中具有可变视图的零镜头对象定制”, *arXiv preprint arXiv:2310.19784*, 2023 年. [118] H. Li, Y. Yang, M. Chang, H. Feng, Z. hai Xu, Q. Li and Y. ting Chen, “Srdiff: 具有扩散概率模型的单图像超分辨率”, *Neurocomputing*, 卷. 479, 第 47-59 页, 2022 年. [119] M. Özbey, S. U. Dar, H. A. Bedel, O. Dalmaz, S. Öztürk, A. Güngör and T. Çukur, “使用对抗性扩散模型进行无监督医学图像翻译”, *arXiv preprint arXiv:2207.08208*, 2022 年. [120] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund and T. B. Schön, “Refusion: 使用潜在空间扩散模型实现大尺寸逼真图像恢复”, *CVPR*, 第 1680-1691 页, 2023 年. [121] Z. Chen, Y. Zhang, D. Liu, B. Xia, J. Gu, L. Kong and X. Yuan, “用于逼真图像去模糊的分层集成扩散模型”, *arXiv preprint arXiv:2305.12966*, 2023 年. [122] F. Guth, S. Coste, V. De Bortoli and S. Mallat, “基于小波分数的生成建模”, 载于 *NeurIPS*, 卷. 35, 2022 年. [123] J. Huang, Y. Liu and S. Chen, “用于高分辨率低光图像增强的引导扩散模型曲线估计”, *arXiv preprint arXiv:2309.14709*, 2023 年. [124] J. Wang, Z. Yue, S. Zhou, K. C. Chan and C. C. Loy, “利用扩散先验实现真实世界图像超分辨率”, *arXiv preprint arXiv:2305.07015*, 2023 年. [125] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao and C. Dong, “Diffbir: 基于生成扩散先验的盲图像恢复”, *arXiv preprint arXiv:2308.15070*, 2023 年. [126] H. Sun, W. Li, J. Liu, H. Chen, R. Pei, X. Zou, Y. Yan and Y. Yang, “Coser: 连接图像和语言以实现认知超分辨率”, *arXiv preprint arXiv:2311.16512*, 2023 年. [127] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao and C. Dong, “扩大规模以实现卓越: 在野外实践模型缩放以实现照片般逼真的图像恢复”, *arXiv preprint arXiv:2401.13627*, 2024 年. [128] J. Song, A. Vahdat, M. Mardani and J. Kautz, “用于逆问题的伪逆引导扩散模型”, *ICLR*, 2022 年. [129] J. Schwab, S. Antholzer and M. Haltmeier, “用于逆问题的深度零空间学习: 收敛分析和速率”, *Inverse Problems*, 第 35 卷, 第 2 期, 第 025008, 2019 年. [130] Y. Wang, Y. Hu, J. Yu and J. Zhang, “基于 Gan 先验的零空间学习实现一致的超分辨率”, 载于 *AAAI*, 第 37 卷, 第 2724-2732 页, 2023 年. [131] G. Kim, T. Kwon and J. C. Ye, “Diffusionclip: 用于稳健图像处理的文本引导扩散模型”, 载于 *CVPR*, 第 2426-2435 页, 2022 年. [132] M. Kwon, J. Jeong and Y. Uh, “扩散模型已经具有语义潜在空间”, 载于 *ICLR*, 2023 年. [133] N. Starodubcev, D. Baranchuk, V. Khruľkov and A. Babenko, “通过无条件扩散模型实现实时文本驱动的图像处理”, *arXiv preprint arXiv:2304.04344*, 2023 年. [134] N. Huang, Y. Zhang, F. Tang, C. Ma, H. Huang, W. Dong and C. Xu, “Diffstyle: 用于文本驱动的图像风格化的可控双重扩散”, *IEEE TNNLS*, 2024 年. [135] Z. Wang, L. Zhao and W. Xing, “StyLEDiffusion: 通过扩散模型实现可控的解缠结风格转移”, 载于 *ICCV*, 第 7677-7689 页, 2023 年. [136] H. Sasaki, C. G. Willcocks and T. P. Breckon, “Unit-ddpm: 具有去噪扩散概率模型的非配对图像转换”, *arXiv preprint arXiv:2104.05358*, 2021 年. [137] S. Xu, Z. Ma, Y. Huang, H. Lee and J. Chai, “Cyclenet: 重新思考用于图像处理的文本引导扩散中的循环一致性”, *NeurIPS*, 2023 年. [138] K. Preechakul, N. Chatthee, S. Widadwongsa and S. Suwajanakorn, “扩散自动编码器: 走向有意义且可解码的表示”, *CVPR*, 第 10619-10629 页, 2022 年. [139] Z. Lu, C. Wu, X. Chen, Y. Wang, L. Bai, Y. Qiao and X. Liu, “分层扩散自动编码器和解缠结图像处理”, in *WACV*, pp. 5374–5383, 2024. [140] M. Zhao, F. Bao, C. Li, and J. Zhu, “Egsde: 通过能量引导的随机微分方程实现非配对图像到图像的转换”, in *NeurIPS*, vol. 35, 第 3609-3623 页, 2022 年. [141] N. Matsunaga, M. Ishii, A. Hayakawa, K. Suzuki and T. Narihira, “使用扩散模型通过像素引导进行细粒度图像编辑”, 载于 *CVPR workshop*, 2023 年. [142] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen and F. Wen, “通过示例绘画: 基于扩散模型的基于样本的图像编辑”, 载于 *CVPR*, 第 18381-18391 页, 2023 年. [143] K. Kim, S. Park, J. Lee and J. Choo, “通过结构感知扩散模型实现基于参考的图像合成与草图”, 载于 *CVPR workshop*, 2023 年. [144] Y. Song, Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim, and D. Aliaga, “Objectstitch: 使用扩散模型进行对象合成”, in *CVPR*, pp. 18310–18319, 2023. [145] X. Zhang, J. Guo, P. Yoo, Y. Matsuo, and Y. Iwasawa, “粘贴, 修复并通过去噪进行协调: 使用预先训练的扩散模型进行主题驱动的图像编辑”, *arXiv preprint arXiv:2306.07596*, 2023. [146] S. Xie, Y. Zhao, Z. Xiao, K. C. Chan, Y. Li, Y. Xu, K. Zhang, and T. Hou, “Dreaminpainter: 文本引导的主题驱动图像修复与扩散模型”, *arXiv preprint arXiv:2312.03771*, 2023. [147] X. 陈, 黄, 刘, 沉, D. Zhao and H. Zhao, “Anydoor: 零样本对象级图像定制”, 载于 *CVPR*, 2024 年. [148] X. Chen and S. Lathuilière, “通过基于扩散的编辑实现面部老化”, 载于 *BMCV*, 2023 年. [149] V. Goel, E. Peruzzo, Y. Jiang, D. Xu, N. Sebe, T. Darrell, Z. Wang and H. Shi, “成对扩散: 使用结构和外观成对扩散模型进行对象级图像编辑”, *arXiv preprint arXiv:2303.17546*, 2023 年. [150] S. Xie, Z. Zhang, Z. Lin, T. Hinz and K. Zhang, “Smartbrush: 使用扩散模型进行文本和形状引导的对象修复”, 载于 *CVPR*, 第 22428-22437 页, 2023 年. [151] Z. Zhang, J. Zheng, Z. Fang and B. A. Plummer, “通过删除图像信息实现文本转图像编辑”, 载于 *WACV*, 第 5232-5241 页, 2024 年. [152] J. Zhuang, Y. Zeng, W. Liu, C. Yuan and K. Chen, “一项任务胜过一个词: 通过任务提示学习高质量多功能图像修复”, *arXiv preprint arXiv:2312.03594*, 2023 年. [153] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet and R. Soricut, et al., “Imagen 编辑器和编辑台: 推进和评估文本引导的图像修复”, 载于 *CVPR*, 第 18359-18369 页, 2023 年. [154] J. Singh, J. Zhang, Q. Liu, C. Smith, Z. Lin and L. Zheng, “Smartmask: 用于细粒度对象插入和布局控制的上下文感知高保真蒙版生成”, 载于 *CVPR*, 2024 年. [155] S. Yang, X. Chen and J. Liao, “Uni-paint: 具有预训练扩散模型的多模态图像修复统一框架”, 载于 *ACM MM*, 第 3190-3199 页, 2023 年. [156] T. Brooks, A. Holynski and A. A. Efros, “Instructpix2pix: 学习遵循图像编辑指令”, 载于 *CVPR*, 第 18392-18402 页, 2023 年. [157] S. Li, C. Chen and H. Lu, “Moecontroller: 基于指令的任意图像处理与混合专家控制器”, *arXiv preprint arXiv:2309.04372*, 2023 年. [158] Q. Guo and T. Lin, “专注于你的指令: 通过注意力调节实现细粒度和多指令图像编辑”, *CVPR*, 2024 年.

- [159] T. Chakrabarty, K. Singh, A. Saakyan 和 S. Muresan, “学习忠实地遵循以对象为中心的图像编辑说明”, 载于 *EMNLP*, 第 9630–9646 页, 2023 年。[160] Z 耿、杨、杭、李、顾、张、包、张、胡、D. Chen 和 G. Baining, “Instructdiffusion: 视觉任务的通用建模界面”, *CVPR*, 2024 年。[161] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh 和 Y. Taigman, “鹈鹕编辑: 通过识别和生成任务进行精确图像编辑”, *arXiv preprint arXiv:2311.10089*, 2023 年。[162] J. Wei, S. Wu, X. Jiang, 和 Y. Wang, “Dialogpaint: 基于对话框的图像编辑模型”, *arXiv preprint arXiv:2303.10073*, 2023。[163] A. B. Yildirim, V. Baday, E. Erdem, A. Erdem 和 A. Dundar, “Inst-inpaint: 指导用扩散模型去除物体”, *arXiv preprint arXiv:2304.03246*, 2023。[164] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon, *et al.*, “Hive: 利用人类反馈进行教学视觉编辑”, *CVPR*, 2024 年。[165] S. Yasheng, Y. Yang, H. Peng, Y. Shen, Y. Yang, H. Hu, L. Qiu 和 H. Koike, “Imagebrush: 学习基于样本的图像处理的视觉上下文指令”, 载于 *NeurIPS*, 2023 年。[166] S. Li, H. Singh 和 A. Grover, “Instructany2pix: 通过多模式指令跟踪实现灵活的视觉编辑”, *arXiv preprint arXiv:2312.06738*, 2023 年。[167] T.-J. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang 和 Z. Gan, “通过多模态大语言模型指导基于指令的图像编辑”, *ICLR*, 2024 年。[168] Y. Huang, L. Xie, X. Wang, Z. Yuan, X. Cun, Y. Ge, J. Zhou, C. Dong, R. Huang, R. Zhang 和 Y. Shan, “Smartedit: 探索复杂性基于指令的图像编辑与多模态大语言模型”, in *CVPR*, 2024。[169] R. Bodur, E. Gundogdu, B. Bhattarai, T.-K. Kim, M. Donoser 和 L. Bazzani, “iedit: 弱监督下的本地化文本引导图像编辑”, *arXiv preprint arXiv:2305.05947*, 2023。[170] Y. Lin, Y.-W. 陈, Y.-H. Tsai, L. Jiang 和 M.-H. Yang, “通过可学习区域进行文本驱动的图像编辑”, 载于 *CVPR*, 2024 年。[171] D. Yue, Q. Guo, M. Ning, J. Cui, Y. Zhu 和 L. Yuan, “Chatface: 通过扩散潜在空间操纵进行聊天引导的真实面部编辑”, *arXiv preprint arXiv:2305.14742*, 2023 年。[172] D. Valevski, M. Kalman, Y. Matias 和 Y. Leviathan, “Unitune: 文本驱动的图像通过在单个图像上微调图像生成模型来编辑”, *ACM TOG*, 2023 年。[173] J. Choi, Y. Choi, Y. Kim, J. Kim 和 S. Yoon, “自定义编辑: 文本-使用定制扩散模型进行引导图像编辑”, 载于 *CVPR workshop*, 2023 年。[174] J. Huang, Y. Liu, J. Qin, 和 S. Chen, “Kv 反演: Kv 嵌入学习用于文本条件真实图像动作编辑”, *arXiv preprint arXiv:2309.16608*, 2023 年。[175] R. Mokady, A. Hertz, K. Aberman, Y. Pritch 和 D. Cohen-Or, “Null - 使用引导扩散模型对真实图像进行文本反转编辑”, 载于 *CVPR*, 第 6038–6047 页, 2023 年。[176] K. Wang, F. Yang, S. Yang, M. A. Butt 和 J. van de Weijer, “动态提示学习: 解决基于文本的图像编辑的交叉注意力泄漏问题”, *NeurIPS*, 2023 年。[177] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang 和 S. Chang, “揭示文本到图像扩散模型中的解缠能力”, *CVPR*, 第 1900-1910 页, 2023 年 6 月。[178] W. Dong, S. Xu, X. Duan 和 S. Han, “使用扩散模型进行文本驱动图像编辑的快速调整反演”, 载于 *ICCV*, 2023 年。[179] S. Li, J. van de Weijer, T. Hu, F. S. Khan, Q. Hou, Y. Wang 和 J. Yang, “StyLEDiffusion: 用于基于文本的编辑的提示嵌入反转”, *arXiv preprint arXiv:2303.15649*, 2023 年。[180] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong 和 C. Xu, “基于反转的扩散模型风格迁移”, 载于 *CVPR*, 第 10146–10156 页, 2023 年 6 月。[181] C. Mou, X. Wang, J. Song, Y. Shan 和 J. Zhang, “DragonDiffusion: 在扩散模型上实现拖拽式操控”, *ICLR*, 2024 年。[182] Y. Shi, C. Xue, J. Pan, W. Zhang, V. Y. Tan 和 S. Bai, “DragDiffusion: 利用扩散模型进行基于点的交互式图像编辑”, 载于 *CVPR*, 2024 年。[183] A. Hertz, K. Aberman 和 D. Cohen-Or, “Delta 去噪分数”, 载于 *ICCV*, 第 2328-2337 页, 2023 年。[184] G. Kwon 和 J. C. Ye, “使用解缠风格和内容表示的基于扩散的图像翻译”, 载于 *ICLR*, 2023 年。[185] H. Nam, G. Kwon, G. Y. Park 和 J. C. Ye, “文本引导的潜在扩散图像编辑的对比去噪分数”, 在 *CVPR* 中, 2024 年。[186] S. Yang, L. Zhang, L. Ma, Y. Liu, J. Fu 和 Y. He, “Magicremover: 使用扩散模型进行无调整文本引导图像修复”, *arXiv preprint arXiv:2310.02848*, 2023。[187] B. Kavar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri 和 M. Irani, “Imagic: 使用扩散模型进行基于文本的真实图像编辑”, 载于 *CVPR*, 第 6007-6017 页, 2023 年 6 月。[188] P. Li, Q. Huang, Y. Ding 和 Z. Li, “LayerDiffusion: 使用扩散模型进行分层控制图像编辑”, in *SIGGRAPH Asia 2023 Technical Communications*, 第 1-4 页, 2023 年。[189] S. 张, S. 肖和 W. 黄, “Forgedit: 通过学习和遗忘进行文本引导的图像编辑”, *arXiv preprint arXiv:2309.10556*, 2023 年。[190] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas 和 J. Ren, “Sine: 使用文本编辑单幅图像-to-image 扩散模型”, 载于 *CVPR*, 第 6027-6037 页, 2023 年。[191] H. Ravi, S. Kelkar, M. Harikumar 和 A. Kale, “Preditor: 利用扩散进行文本引导图像编辑先前的”, *arXiv preprint arXiv:2302.07979*, 2023 年。[192] Y. Lin, S. Zhang, X. Yang, X. Wang, 和 Y. Shi, “具有丰富提示的零样本图像翻译扩散模型的再生学习”, *arXiv preprint arXiv:2305.04651*, 2023 年。[193] S. Kim, W. Jang, H. Kim, J. Kim, Y. Choi, S. Kim 和 G. Lee, “使用最少的文本输入进行用户友好的图像编辑: 利用字幕和注入技术”, *arXiv preprint arXiv:2306.02717*, 2023 年。[194] Q. Wang, B. Zhang, M. Birsak 和 P. Wonka, “Instructedit: 使用用户指令改进基于扩散的图像编辑的自动蒙版”, *arXiv preprint arXiv:2305.18047*, 2023 年。[195] A. Elarabawy, H. Kamath 和 S. Denton, “直接反演: 无优化文本驱动的真实使用扩散模型进行图像编辑”, *arXiv preprint arXiv:2211.07825*, 2022 年。[196] I. Huberman-Spiegelglas, V. Kulikov 和 T. Michaeli, “易于编辑的 ddpm 噪声空间: 反演和操作”, *CVPR*, 2024 年。[197] 聂志军、郭华安、陆聪、周颖、郑聪、李聪, “随机性的祝福: Sde 在一般基于扩散的图像编辑中胜过 ode”, 载于 *ICLR*, 2024 年。[198] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting 和 A. Passos, “Ledits++: 使用文本到图像模型进行无限图像编辑”, 载于 *CVPR*, 2024 年。[199] S. Chen 和 J. Huang, “Fec: 三种无需微调的增强真实图像编辑一致性的方法”, *arXiv preprint arXiv:2309.14934*, 2023 年。[200] K. Joseph, P. Udhayan, T. Shukla, A. Agarwal, S. Karanam, K. Goswami 和 B. V. Srinivasan, “迭代多-使用扩散模型进行粒度图像编辑”, 载于 *WACV*, 第 8107-8116 页, 2024。[201] D. Miyake, A. Iohara, Y. Saito 和 T. Tanaka, “负提示反演: 使用文本引导扩散模型进行编辑的快速图像反演”, *arXiv preprint arXiv:2305.16807*, 2023 年。[202] L. Han, S. Wen, Q. Chen, Z. Zhang, K. Song, M. Ren, R. Gao, A. Stathopoulos, X. He, Y. Chen, *et al.*, “Proxedit: 改进调音基于近端引导的自由真实图像编辑”, 载于 *WACV*, 第 4291-4301 页, 2024 年。[203] J. Zhao, H. Zheng, C. Wang, L. Lan, W. Huang, 和 W. Yang, “传播模型中的空文本指导其实是一位卡通风格的创造者”, 载于 *ACM MM*, 第 5143-5152 页, 2023。[204] B. Wallace, A. Gokul 和 N. Naik, “Edict: 通过耦合变换实现精确扩散反演”, *CVPR*, 第 22532-22541 页, 2023 年。[205] Z. Pan, R. Gherardi, X. Xie 和 S. Huang, “利用加速迭代扩散反演实现有效的实像编辑”, 载于 *ICCV*, 第 15912–15921 页, 第 2023 页。[206] C. H. Wu 和 F. De la Torre, “用于零样本图像编辑和指导的随机扩散模型的潜在空间”, 载于 *ICCV*, 第 7378-7387 页, 2023 年。[207] J. Jeong, M. Kwon 和 Y. Uh, “无需训练的使用 h 空间进行内容注入在扩散模型中”, 载于 *WACV*, 第 5151-5161 页, 2024 年。[208] B. Meiri, D. Samuel, N. Darshan, G. Chechik, S. Avidan 和 R. Ben-Ari, “固定-point inversion for text-to-image transmission models”, *arXiv preprint arXiv:2312.12540*, 2023 年。[209] X. Duan, S. Cui, G. Kang, B. Zhang, Z. Fei, M. Fan 和 J. Huang, “无需调整的反转增强控制, 实现一致的图像编辑”, *arXiv preprint arXiv:2312.14611*, 2023 年。[210] P. Gholami 和 R. Xiao, “扩散画笔: 一种基于潜在扩散模型的人工智能编辑工具”, 生成的图像”, *arXiv preprint arXiv:2306.00219*, 2023 年。[211] D. Epstein, A. Jabri, B. Poole, A. A. Efros 和 A. Holynski, “扩散自引导可控图像生成”, 载于 *NeurIPS*, 2023 年。[212] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman 和 Y. Pritch, 以及 D. Cohen-Or, “通过交叉注意力控制的提示到提示图像编辑”, *ICLR*, 2023 年。

- [213] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu and J.-Y. Zhu, “零样本图像到图像翻译”, 载于 *ACM SIGGRAPH*, 第 1-11 页, 2023 年。[214] N. Tumanyan, M. Geyer, S. Bagon 和 T. Dekel, “即插即用扩散特征用于文本驱动的图像到图像翻译”, 载于 *CVPR*, 第 1921-1930 页, 2023 年。[215] S. Lu, Y. Liu 和 A. W.-K. Kong, “Tf-icon: 基于扩散的无训练跨域图像合成”, 载于 *ICCV*, 第 2294-2305 页, 2023 年。[216] O. Patashnik, D. Garibi, I. Azuri, H. Averbuch-Elor 和 D. Cohen-Or, “使用文本到图像扩散模型定位对象级形状变化”, 载于 *ICCV*, 第 23051-23061 页, 2023 年。[217] H. Lee, M. Kang 和 B. Han, “文本驱动的图像到图像翻译的条件分数指导”, 载于 *NeurIPS*, 2023 年。[218] G. Y. Park, J. Kim, B. Kim, S. W. Lee 和 J. C. Ye, “基于能量的交叉注意用于文本到图像扩散模型中的贝叶斯上下文更新”, 载于 *NeurIPS*, 2023 年。[219] D. H. Park, G. Luo, C. Toste, S. Azadi, X. Liu, M. Karalashvili, A. Rohrbach 和 T. Darrell, “具有内外注意的形状引导扩散”, 载于 *WACV*, 第 4198-4207 页, 2024 年。[220] H. Manukyan, A. Sargsyan, B. Atanyan, Z. Wang, S. Navasardyan 和 H. Shi, “Hd-painter: 使用扩散模型进行高分辨率和及时忠实的文本引导图像修复”, *arXiv preprint arXiv:2312.14091*, 2023 年。[221] Z. Yu, H. Li, F. Fu, X. Miao 和 B. Cui, “Fisedit: 通过支持缓存的稀疏扩散推理加速文本到图像编辑”, 载于 *AAAI*, 2024 年。[222] O. Avrahami, O. Fried 和 D. Lischinski, “混合潜在扩散”, *ACM TOG*, 第 42 卷, 第 4 期, 第 1-11 页, 2023 年。[223] W. Huang, S. Tu 和 L. Xu, “Pfb-diff: 用于文本驱动图像编辑的渐进式特征混合扩散”, *arXiv preprint arXiv:2306.16894*, 2023 年。[224] G. Coiairon, J. Verbeek, H. Schwenk 和 M. Cord, “Diffedit: 基于扩散的带蒙版引导的语义图像编辑”, *ICLR*, 2023 年。[225] N. Huang, F. Tang, W. Dong, T.-Y. Lee 和 C. Xu, “区域感知扩散用于零样本文本驱动的图像编辑”, *arXiv preprint arXiv:2302.11797*, 2023 年。[226] Z. Liu, F. Zhang, J. He, J. Wang, Z. Wang 和 L. Cheng, “文本引导的无蒙版局部图像修饰”, *ICME*, 第 2783-2788 页, 2023 年。[227] E. Levin 和 O. Fried, “差异扩散: 赋予每个像素强度”, *arXiv preprint arXiv:2306.00950*, 2023 年。[228] A. Mirzaei, T. Aumentado-Armstrong, M. A. Brubaker, J. Kelly, A. Levinstein, K. G. Derpanis 和 I. Gilitschenski, “注意你的步骤: 通过文本指令编辑本地图像和场景”, *arXiv preprint arXiv:2308.08947*, 2023 年。[229] O. Avrahami, D. Lischinski 和 O. Fried, “混合扩散用于文本驱动的自然图像编辑”, 载于 *CVPR*, 第 18208-18218 页, 2022 年。[230] S. Li, B. Zeng, Y. Feng, S. Gao, X. Liu, J. Liu, L. Lin, X. Tang, Y. Hu, J. Liu 和 B. Zhang, “区域: 零样本指令引导的局部编辑”, 载于 *CVPR*, 2024 年。[231] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng 和 Z. Chen, “修复一切: 分割一切与图像修复相遇”, *arXiv preprint arXiv:2304.06790*, 2023 年。[232] M. Brack, P. Schramowski, F. Friedrich, D. Hintersdorf 和 K. Kersting, “稳定的艺术家: 扩散潜在空间中的转向语义”, *arXiv preprint arXiv:2212.06013*, 2022 年。[233] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski 和 K. Kersting, “Sega: 使用语义指导指导文本到图像模型”, *NeurIPS*, 2023 年。[234] L. Tsaban 和 A. Passos, “Ledits: 使用 ddpm 反转和语义指导进行真实图像编辑”, *arXiv preprint arXiv:2307.00522*, 2023 年。[235] Z. Yang, D. Gui, W. Wang, H. Chen, B. Zhuang 和 C. Shen, “用于图像编辑的对象感知反转和重组”, *ICLR*, 2024 年。[236] T. Karras, T. Aila, S. Laine 和 J. Lehtinen, “逐步发展 gans 以提高质量、稳定性和变化”, 收录于 *ICLR*, 2018 年。[237] Y. Choi, Y. Uh, J. Yoo 和 J.-W. Ha, “Stargan v2: 适用于多个领域的多样化图像合成”, 收录于 *CVPR*, 第 8188-8197 页, 2020 年。[238] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser 和 J. Xiao, “Lsun: 使用深度学习与人类参与构建大规模图像数据集”, *arXiv preprint arXiv:1506.03365*, 2015 年。[239] W. 志愿者团队, “Wiki art 数据集”。<https://www.wikiart.org/>。[240] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik 和 D. Cohen-Or, 《Stylegan-nada: 图像生成器的剪辑引导域自适应》, *ACM TOG*, 第 41 卷, 第 4 期, 第 1-13 页, 2022 年。[241] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or 和 D. Lischinski, “Styleclip: 文本驱动的 Stylegan 图像处理”, 载于 *ICCV*, 第 2085-2094 页, 2021 年。[242] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “从自然语言监督中学习可转移视觉模型”, 载于 *ICML*, 第 8748-8763 页, 2021 年。[243] Y. Shen, J. Gu, X. Tang 和 B. Zhou, “解释 GAN 的潜在空间以进行语义人脸编辑”, 载于 *CVPR*, 第 9243-9252 页, 2020 年。[244] R. Abdal, Y. Qin 和 P. Wonka, “Image2stylegan++: 如何编辑嵌入的图像?”, 载于 *CVPR*, 第 8296-8305 页, 2020 年。[245] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry 和 A. Askell, “语言模型是小样本学习者”, 载于 *NeurIPS*, 卷 33, 第 1877-1901 页, 2020 年。[246] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar 和 R. Girshick, “Segment anything”, 载于 *ICCV*, 第 4015-4026 页, 2023 年。[247] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, “Chain-of-thought prompting elicits logic in large language models”, 载于 *NeurIPS*, 第 35, 第 24824-24837 页, 2022 年。[248] J. Li, D. Li, S. Savarese 和 S. Hoi, “Blip-2: 使用冻结图像编码器和大型语言模型进行引导语言图像预训练”, *ICML*, 2023 年。[249] K. Zhang, L. Mo, W. Chen, H. Sun 和 Y. Su, “Magicbrush: 用于指令引导图像编辑的手动注释数据集”, *arXiv preprint arXiv:2306.10012*, 2023 年。[250] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. B. Ashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, “Llama 2: 开放基础和微调聊天模型”, *arXiv preprint arXiv:2307.09288*, 2023 年。[251] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi 和 H. Hajishirzi, “自我指导: 将语言模型与自生成指令对齐”, *arXiv preprint arXiv:2212.10560*, 2022 年。[252] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, *et al.*, “构建开放域聊天机器人的秘诀”, 载于 *ACL*, 第 300-325 页, 2021 年。[253] D. A. Hudson 和 C. D. Manning, “Gqa: 用于现实世界视觉推理和组合问答的新数据集”, 载于 *CVPR*, 第 6700-6709 页, 2019 年。[254] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo 和 R. Girshick, “Detectron2”。<https://github.com/facebookresearch/detectron2>, 2019 年。[255] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl 和 I. Misra, “使用图像级监督检测两万个类别”, 载于 *ECCV*, 2022 年。[256] Y. Zeng, Z. Lin, H. Lu 和 V. M. Patel, “Cr-fill: 带有辅助上下文重建的生成性图像填充”, 载于 *ICCV*, 第 14164-14173 页, 2021 年。[257] R. Giridhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin 和 I. Misra, “Imagebind: 一个嵌入空间将它们全部绑定”, 载于 *CVPR*, 第 15180-15190, 2023。[258] W.-L. Jiang, Z. Li, Z. Lin, Y. Shen, Z. Wu, H. Zhang, L. Cheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, *et al.*, “Vicuna: 开源聊天机器人以 90%* chatgpt 质量打动 gpt-4”, *See https://vicuna.lmsys.org (accessed 14 April 2023)*, 2023。[259] H. Liu, C. Li, Q. Wu 和 Y. J. Lee, “视觉指令调整”, *NeurIPS*, 卷. 36, 2023 年。[260] T. Lüddecke 和 A. Ecker, “使用文本和图像提示进行图像分割”, *CVPR*, 第 7086-7096 页, 2022 年。[261] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni 和 H.-Y. Shum, “Dino: 带有改进的去噪锚框的 Detr 用于端到端物体检测”, 载于 *ICLR*, 2023 年。[262] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka 和 C. Theobalt, “Drag your gan: 在生成图像流形上进行交互式点击操作”, 载于 *ACM SIGGRAPH*, 第 1-11 页, 2023 年。[263] J. Li, D. Li, C. Xiong 和 S. Hoi, “Blip: 引导语言-图像预训练以实现统一的视觉-语言理解和生成”, 载于 *ICML*, 第 1288-12900 页, 2022 年。[264] M. A. Chan, S. I. Young 和 C. A. Metzler, “Sud²: 用于图像重建的去噪扩散模型”, *arXiv preprint, arXiv:2303.09642*, 2023 年。[265] S. I. Young, A. V. Dalca, E. Ferrante, P. Golland, C. A. Metzler, B. Fischl 和 J. E. Iglesias, “通过去噪进行监督”, *IEEE TPAMI*, 2023 年。[266] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte 和 L. Van Gool, “重新绘制: 使用去噪扩散概率模型进行修复”, *CVPR*, 第 11461-11471 页, 2022 年。

- [267] A. Grechka, G. Couairon 和 M. Cord, “Gradpaint: 基于扩散模型的梯度引导修复”, *arXiv preprint arXiv:2309.09614*, 2023 年。[268] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky 和 J. C. Ye, “用于一般噪声逆问题的扩散后验采样”, 载于 *ICLR*, 2023 年。[269] J. Song, A. Vahdat, M. Mardani 和 J. Kautz, “用于逆问题的伪逆引导扩散模型”, 载于 *ICLR*, 2023 年。[270] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang 和 B. Dai, “用于统一图像恢复和增强的生成扩散先验”, 载于 *CVPR*, 第 9935–9946 页, 2023 年。[271] G. Zhang, J. Ji, Y. Zhang, M. Yu, T. Jaakkola 和 S. Chang, “利用去噪扩散隐式模型实现相干图像修复”, 载于 *ICML*, 第 41164–41193 页, 2023 年。[272] Z. Fabian, B. Tinaz 和 M. Soltanolkotabi, “Diracdiffusion: 确保数据一致性的去噪和增量重建”, *arXiv preprint arXiv:2303.14353*, 2023 年。[273] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov 和 L.-C. Chen, “Mobilenetv2: 倒置残差和线性瓶颈”, *CVPR*, 第 4510–4520 页, 2018 年。[274] J. Xu, S. Motamed, P. Vaddamanu, C. H. Wu, C. Haene, J.-C. Bazin 和 F. De la Torre, “通过并行视觉注意的扩散模型进行个性化人脸修复”, 载于 *WACV*, 第 5432–5442 页, 2024 年。[275] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park 和 V. Lempitsky, “通过傅里叶卷积进行分辨率稳健的大蒙版修复”, 载于 *WACV*, 第 2149–2159 页, 2022 年。[276] S. Basu, M. Saberi, S. Bhardwaj, A. M. Chegini, D. Massiceti, M. Sanjabi, S. X. Hu 和 S. Feizi, “Editval: 基于扩散的文本引导图像编辑方法基准测试”, *arXiv preprint arXiv:2310.02426*, 2023 年。[277] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár 和 C. L. Zitnick, “Microsoft coco: 上下文中的常见对象”, 载于 *ECCV*, 第 740–755 页, 2014 年。[278] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee 和 Y. Guo, *et al.*, “通过更好的字幕改进图像生成”, *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 第 2 卷, 第 2 号 3, 2023 年。[279] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras 和 Y. Choi, “Clipscore: 一种无参考的图像字幕评估指标”, 载于 *EMNLP*, 第 7514–7528 页, 2021 年。[280] A. Sauer, D. Lorenz, A. Blattmann 和 R. Rombach, “对抗性扩散蒸馏”, *arXiv preprint arXiv:2311.17042*, 2023 年。[281] Z. Geng, A. Pokle 和 J. Z. Kolter, “通过深度平衡模型实现一步步扩散蒸馏”, 第 36, 2023 年。[282] Y. Song, P. Dhariwal, M. Chen 和 I. Sutskever, “一致性模型”, *ICML*, 2023 年。[283] S. Luo, Y. Tan, L. Huang, J. Li 和 H. Zhao, “潜在一致性模型: 通过几步推理合成高分辨率图像”, *arXiv preprint arXiv:2310.04378*, 2023 年。[284] Y. Zhao, Y. Xu, Z. Xiao 和 T. Hou, “Mobilediffusion: 移动设备上亚秒级文本到图像生成”, *arXiv preprint arXiv:2311.16567*, 2023 年。[285] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov 和 J. Ren, “Snapfusion: 两秒内移动设备上的文本到图像扩散模型”, *NeurIPS*, 卷 36, 2023 年。[286] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang 和 W. Chen, “Lora: 大型语言模型的低秩适应”, *arXiv preprint arXiv:2106.09685*, 2021 年。[287] P. Schramowski, M. Brack, B. Deiseroth 和 K. Kersting, “安全的潜在扩散: 减轻扩散模型中的不适当退化”, *CVPR*, 第 22522–22531 页, 2023 年。[288] R. Pandey, S. O. Escolano, C. Legendre, C. Haene, S. Bouaziz, C. Rhemann, P. Debevec 和 S. Fanello, “全面重新照明: 学习重新照明肖像以替换背景”, *ACM TOG*, 第 40, 没有, 4, 第 1–21 页, 2021 年。[289] P. Ponglertnapakorn, N. Tritrong 和 S. Suwajanakorn, “Difareli: 扩散人脸重新照明”, *arXiv preprint arXiv:2304.09479*, 2023 年。[290] M. Ren, W. Xiong, J. S. Yoon, Z. Shu, J. Zhang, H. Jung, G. Gerig 和 H. Zhang, “重新照明的协调: 照明感知的肖像背景替换”, *arXiv preprint arXiv:2312.06886*, 2023 年。[291] J. Xiang, J. Yang, B. Huang 和 X. Tong, “使用 2d 扩散模型的 3d 感知图像生成”, *arXiv preprint arXiv:2303.17905*, 2023 年。[292] J. Ackermann 和 M. Li, “通过多阶段混合扩散实现高分辨率图像编辑”, *arXiv preprint arXiv:2210.12965*, 2022 年。[293] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler 和 S. Hochreiter, “通过两个时间尺度更新规则训练的 Gans 收敛到局部纳什均衡”, 载于 *NeurIPS*, 第 30 卷, 2017 年。[294] M. Bińkowski, D.J. Sutherland, M. Arbel 和 A. Gretton, “揭开 mm d gans 的神秘面纱”, *arXiv preprint arXiv:1801.01401*, 2018 年。[295] R. Zhang, P. Isola, A. A. Efros, E. Shechtman 和 O. Wang, “深度特征作为感知指标的不合理有效性”, *CVPR*, 第 586–595 页, 2018 年。[296] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel 和 P. Isola, “Dreamsim: 使用合成数据学习人类视觉相似性的新维度”, *arXiv preprint arXiv:2306.09344*, 2023 年。[297] K. Kotar, S. Tian, H.-X. Yu, D. Yamins 和 J. Wu, “这些是同一个苹果吗? 根据对象内在特性比较图像”, 载于 *NeurIPS*, 第 36 卷, 2023 年。