arXiv:2305.03382v3 [cs.CV] 9 Oct 2024

通过扩散模型中的初始图像编辑实现引导式图像合成

Jiafeng Mao
东京大学
日本东京
mao@hal.t.u-tokyo.ac.jp

Xueting Wang
CyberAgent, Inc.
日本东京
wang_xueting@cyberagent.co.jp

Kiyoharu Aizawa
东京大学
日本东京
aizawa@hal.t.u-tokyo.ac.jp

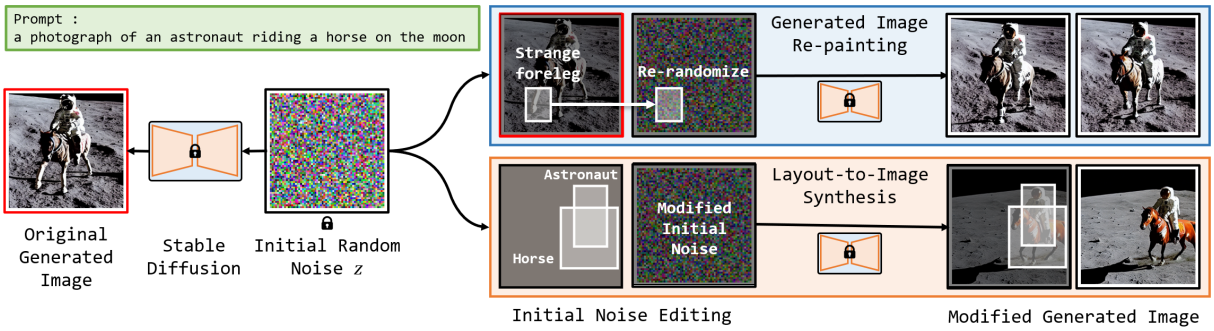

图 1：在本研究中，我们探究初始图像对图像生成的影响，并提出一种通过操控初始随机噪声来控制生成过程的新方向。我们展示了两项研究成果的直接应用：生成图像重绘和布局到图像合成。生成图像重绘允许用户修改生成图像的一部分，同时保留其大部分内容。布局到图像合成则要求在用户指定的区域内生成物体。

摘要

扩散模型能够通过对纯高斯噪声图像进行去噪来生成高质量图像。尽管以往的研究主要集中于通过调整去噪过程来提高图像生成的可控性，但我们提出了一种通过操纵初始噪声来控制生成图像的新颖方向。通过在稳定扩散模型上进行实验，我们发现初始潜在图像中的像素块倾向于生成特定内容，并且修改这些像素块能够显著影响生成图像。具体而言，我们表明，修改初始图像的一部分会影响生成图像中对应的区域，同时不影响其他区域，这对于重绘任务非常有用。此外，我们发现，像素块的生成偏好主要由其数值决定，而非其位置。通过将倾向于生成用户期望内容的像素块移动到用户指定的区域，我们的方法在布局到图像生成任务中实现了最先进的性能。我们的研究结果凸显了初始图像操纵在控制生成图像方面的灵活性和强大能力。项目页面：https://ut-mao.github.io/swap.github.io

CCS 概念

• **计算方法 → 计算机视觉**。

特此授权免费制作本作品的全部或部分数字版或纸质副本，供个人或课堂使用，但前提是不得出于营利或商业目的制作或分发副本，且副本首页必须附有本声明及完整引文。对于本作品中由作者以外的其他方拥有的组件，其版权必须得到尊重。经注明出处后，允许摘编。若要以其他方式复制、重新出版、发布到服务器或重新分发至列表，则须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。*MM '23，2023年10月29日至11月3日*，加拿大安大略省渥太华。© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 979-8-4007-0108-5/23/10...$15.00
https://doi.org/10.1145/3581783.3612191

关键词

文本到图像、扩散模型、细粒度控制、布局到图像

ACM 引用格式：

Jiafeng Mao、Xueting Wang 和 Kiyoharu Aizawa。2023。通过扩散模型中的初始图像编辑实现引导图像合成。载于*第31届 ACM 多媒体国际会议论文集（MM'23），2023年10月29日至11月3日*，加拿大安大略省渥太华。ACM，纽约州纽约市，美国，共9页。https://doi.org/10.1145/3581783.3612191

1 引言

图像生成是计算机视觉领域的重要研究方向。扩散模型 [9] 在通过逐步对纯高斯噪声图像进行去噪来生成高质量图像方面展现出巨大潜力。此前，该领域的研究 [4, 13, 20, 23–25] 主要集中于使用提示词引导图像生成。提示词通常是描述生成图像所需内容的一句话。随后，模型会尝试生成与提示词描述相符的图像。然而，基于提示词的文本到图像生成具有很大的不确定性，而且经常失败。尽管一些方法 [2, 5, 7, 12, 16, 21] 尝试通过改进去噪过程或探索更有效的提示词使用方式，实现更可控的生成过程，但在实际应用中，用户通常需要使用同一提示词随机生成几十张甚至上百张不同的图像，才能获得满意的结果。

在本文 [27]中，我们探究生成有时会失败的关键原因，并进一步将所得结论用于提出一种控制生成内容的新方向。我们注意到，目前广泛使用的采样方法，即 DDIM（去噪扩散隐式模型 [27]）和 PLMS（伪线性多步方法 [15]），都是确定性的。给定一个扩散模型 θ、提示词

[1]本研究是在第一作者于 CyberAgent, Inc. 实习期间完成的。

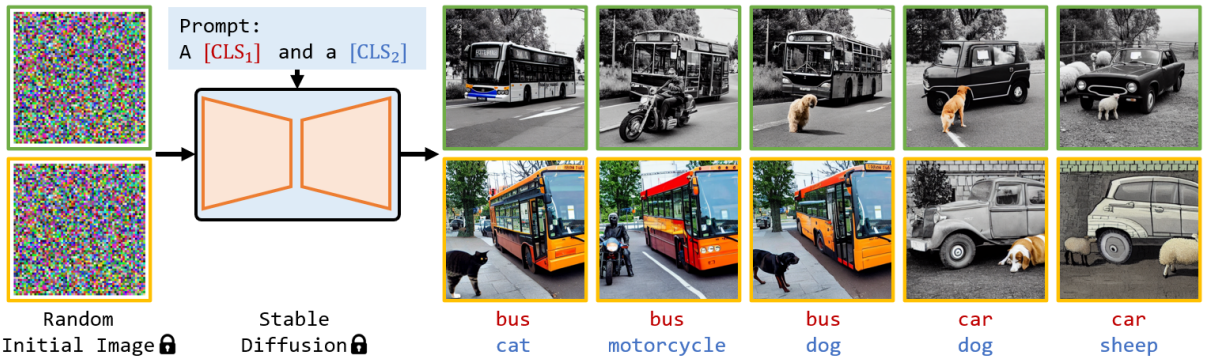

图 2：我们的分析实验聚焦于随机初始噪声图像的生成倾向。我们展示了使用两个不同初始图像生成的结果样本，这两个初始图像均从高斯分布中随机采样。边框颜色相同的图像由同一个初始图像生成。当初始图像相同时，不同提示词下的相同类别在位置和视觉外观上高度相似。当随机采样的初始图像不同时，即使使用相同的提示词，生成结果在布局和细节上也大不相同。

p和一张初始噪声图像 x_T，生成图像 x_0是固定的，可表示为 $x_0 = f_\theta(x_T,\ p)$。在提示词 p和模型参数 θ相同的情况下，最终生成图像 x_0应完全由随机采样得到的初始图像 x_T决定。尽管初始噪声图像处于如此关键的位置，却鲜少受到关注。这一结论促使我们进一步分析初始图像对生成结果的影响。

具体而言，我们对稳定扩散的实验表明，初始噪声图像倾向于生成特定内容，我们在本文中将其称为**生成倾向**。具体来说，一旦初始图像确定，即使使用不同的提示词，生成图像在布局、背景和风格等方面也往往表现出显著的相似性。此外，我们发现初始图像中各区域的生成倾向相对独立。在提示词固定的情况下，修改初始图像中单个区域的像素，可以改变生成图像中的对应区域，同时保留其他区域的内容。这一发现可应用于重绘任务：用户对生成图像的大部分内容感到满意，但希望更改其中一小部分不合适的内容（图 1 上排）。我们还发现，初始图像中各区域的生成倾向主要与其像素值有关，而非其空间位置。倾向于生成特定内容的区域，在初始图像中移动到其他空间位置后，仍可能生成相同内容。因此，只需将所有倾向于生成特定内容的区域移动到指定区域，就能使扩散模型在该区域生成相应内容。这一发现可以自然地应用于布局到图像合成：该任务要求在用户指定的区域内生成提示词中提及的物体（图 1 下排）。单独使用时，我们修改初始图像的方法可达到与最先进的方法相当的性能；将其与这些方法结合，还能进一步提升对物体位置的控制能力。

通过将初始图像用作控制机制，我们能够实现对生成图像的细粒度控制，而不是仅依赖提示词或去噪过程。我们的研究结果表明，初始图像操控是控制扩散模型图像生成的一种灵活而强大的方式，并为图像生成领域的研究开辟了新的途径。总结本文的贡献如下：

- 我们的实验结果表明，初始噪声图像在生成某些内容时表现出明显的偏好，这些偏好主要由其像素值决定。换言之，初始图像会显著影响生成结果。
- 我们找到了文生图有时会失败的一个关键原因。
- 我们提出了一种控制生成的新方向，即将初始图像用作控制机制。我们的方法能够重绘生成图像，并在免训练的布局到图像生成任务中实现最先进的性能。

2 相关工作

2.1 扩散模型

近期关于扩散模型的研究 [9, 19, 27] 已证明，该模型能够生成与真实图像相媲美的样本。DDPM（去噪扩散概率模型 [9]）从纯高斯噪声开始，不断重复预测下一个分布并随机采样新样本的过程，最终生成与真实图像相似的高质量样本。然而，DDPM 需要数百到数千次迭代才能生成最终样本。一些先前的研究通过调整方差方案（改进的 DDPM [19]）或去噪方程（DDIM [27]），成功加速了 DDPM。DDIM 基于非马尔可夫过程，通过在每次迭代中执行多个步骤，并以确定性生成取代随机生成，加速了去噪过程。Liu 等人 [15]将 DDIM 扩展为 PNDM，并提出了伪线性多步方法（PLMS），在生成质量相近的情况下实现了最快的生成速度。以往研究提出了多种控制生成内容的方法。分类器引导 [3] 根据分类器的梯度，在朝向对应类别高分的方向更新图像的同时对其进行去噪。无分类器引导 [10] 无需分类器即可实现相同效果。目前大多数方法 [4, 6, 13, 24] 采用提示词引导，并通过 CLIP [22] 计算更新方向来引导去噪。稳定扩散 [25]在潜在空间中进行去噪，然后将潜变量解码为图像，

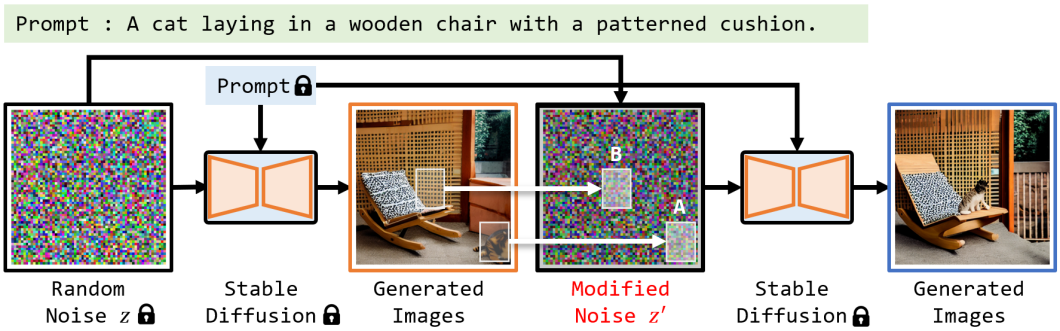

图 3：我们对初始噪声图像编辑进行的实验。我们首先使用从高斯噪声中采样得到的随机噪声初始图像进行生成。尽管第一张生成图像中同时生成了“猫”和“椅子”，但猫的位置并不符合提示词的要求。由于生成猫的区域（边界框 A）倾向于生成“猫”，但该区域并非我们期望的位置，因此我们对初始图像中对应区域的像素值进行重新随机化。此外，我们还对椅子附近的区域（边界框 B）进行重新随机化，因为该区域并不倾向于生成“猫”。经过几次尝试后，使用修改后的初始图像和相同的提示词生成了右侧的改进图像。

这是它通过更多注意的机制 [28] 扩展了 U-Net [26] 骨干网络，从而支持文本等条件形式的灵活输入。

2.2 细粒度控制

仅依靠文本引导进行生成，有时难以取得令人满意的结果，因为细粒度控制（例如物体的大小和位置、图像风格）可能难以通过文本表达，而且生成结果也未必总能符合文本描述。另一方面，扩散模型生成的图像具有很高的不确定性。即使使用相同的提示词，每次生成的结果也可能有很大差异。因此，一些研究 [1, 12, 16, 21]尝试实现更可控的生成过程，以提高扩散模型的实用性。Composable Diffusion [16] 旨在通过接收多个文本提示词，实现对图像生成的灵活控制。具体而言，它将与这些提示词对应的引导向量相加，并将其作为去噪方向。T2I-Adapter [18] 使用适配器提供额外引导，从而丰富生成结果在颜色和结构方面的控制与编辑效果。一些方法 [11, 29, 32] 尝试通过改进模型参数实现更好的控制。PITI [31] 提出通过缩小其他类型条件的特征与文本条件之间的距离来提供结构引导。Sketch-Guided [29] 提出利用目标草图与中间结果之间的相似度梯度，约束最终结果的结构。研究发现，交叉注意力图与生成图像的布局和结构高度对应 [5, 7],，因此有若干研究尝试通过操控注意力图实现细粒度控制。结构化扩散引导 [5]提出预先分析文本的语义结构，并替换特定词语对应的注意力图，以获得更好的 CLIP 嵌入。总体而言，这些工作着重通过改进生成过程和提示词的使用来提升扩散模型的可控性，却忽视了初始图像的作用，而这正是本文关注的重点。

2.3 布局到图像合成

布局到图像合成尝试通过引入逐物体引导来实现更可控的生成。除了提示词之外，布局到图像合成还会采用以下信息：提示词中描述的物体的位置和大小

提示词中描述的物体 生成模型需要生成图像，.使每个物体都位于用户指定的位置。一些研究 [2, 17] 提出在注意力图中添加掩码，以提高在指定区域生成指定物体的概率，从而实现布局到图像合成。这些方法的根本局限在于，它们强行向用户指定的区域添加特征，完全忽略了初始图像的生成倾向。相比之下，我们将初始图像用作控制机制，引导扩散模型主动在预期区域生成用户指定的内容。

3 初始图像编辑

本节将介绍我们的发现和方案。我们在第 3.1 节中展示了相关实验，这些实验揭示了初始图像在用于图像生成的扩散模型中的作用，并表明初始图像的不同区域具有生成特定内容的内在倾向。我们在第 3.2 节分析这一现象，并阐释文生图有时会失败的关键原因。我们还尝试操控初始图像来编辑生成图像，使这一发现适用于图像重绘任务（第 3.3 节）和布局到图像合成任务（第 3.4 节）。

3.1 初始图像的生成倾向

我们首先进行实验，以检查初始图像的生成倾向。具体而言，我们创建一个类别列表，然后从中选择两个不同的类别，CLS_1 和 CLS_2，来构建用于生成的提示词。提示词采用如下形式：［一个 CLS_1 和一个 CLS_2］（例如［一只狗和一辆汽车］）。如图 2 所示，我们使用固定的随机初始图像和具有固定参数的模型来生成每个构建的提示词，并观察同一初始化图像在不同提示词引导下的生成效果。当不同提示词中提及同一物体时，该物体很可能会出现在相同位置，并具有非常相似的视觉外观。例如，［一辆公交车和一只猫］,［一辆公交车和一辆摩托车］,［一辆公交车和一只狗］都包含“公交车”一词，使用同一初始图像但不同提示词生成的公交车看起来很相似。然而，使用不同的初始噪声图像（底行）时，生成图像的构成可能会有很大差异，即使在即刻

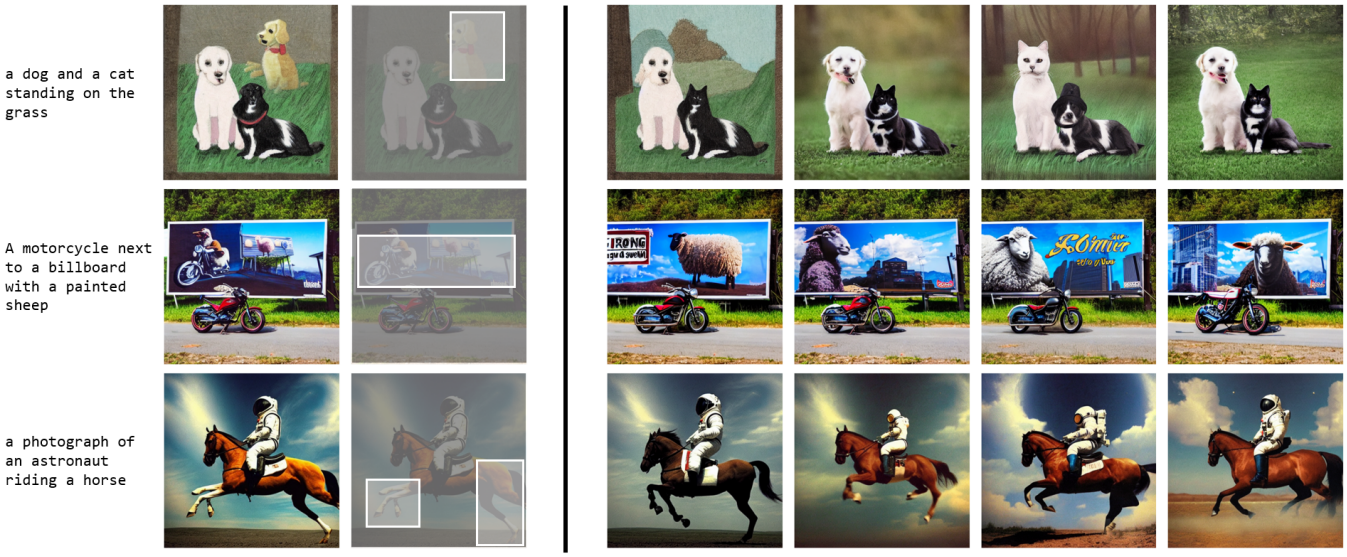

图 4：重新绘制实验样本。左侧图像由随机初始噪声生成，右侧图像由部分重新采样的初始噪声生成。第一个提示词涉及两个物体，但生成了三个物体。重新采样初始图像中与第三个物体对应区域的像素值后，第三个物体被移除了。第二次生成未能生成广告牌的内容（应该是一只羊）。重新采样广告牌对应区域的像素值后，我们得到了布局和风格相同、且广告牌内容生成正确的图像。第三张生成图像中的马有三条前腿，后腿也不自然。重新采样这两个部分后，我们得到了不同且更好的马，同时保持图像整体不变。平均需要尝试 4–5 次才能获得每个修改后的样本。

相同的提示词。基于这一实验，我们发现初始化图像本身携带某些生成倾向。一旦确定初始化图像，其生成倾向就会固定下来，并与提示词共同影响最终生成的图像。

3.2 分析与解释

本节将分析并解释初始图像的生成倾向。

稳定扩散中的确定性去噪 稳定扩散 [25]可以解释为一系列去噪自编码器 $\epsilon_\theta(z_t,\ t)$、$t = 1,\ 2,\ \ldots\ldots,\ T$。这些自编码器经过训练，能够预测输入 z_t 上添加的噪声，其中 z_t是加噪后的 z_0。为了生成真实图像的潜变量，稳定扩散模型从纯高斯噪声 z_T 开始，迭代地对潜在图像进行去噪。稳定扩散通过交叉注意力机制 [28] 增强 U-Net [26] 骨干网络，从而支持文本等灵活的条件形式。对于文生图任务，模型使用 CLIP 将条件 p 投影到中间表示 $\tau_\theta(p)$，随后通过交叉注意力层将其映射到 U-Net 的中间层，如下所示，

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V, \tag{1}$$

其中 $Q = W^{(i)}_Q \cdot z_t,\ K = W^{(i)}_K \cdot \tau_\theta(p)\ V = W^{(i)}_V \cdot \tau_\theta(p)$。因此，给定参数为 $(W,\ \theta)$ 的训练模型，交叉注意力图 $\text{softmax}(\frac{QK^T}{\sqrt{d}})$ 的计算完全取决于初始隐变量 z_t和用户指定的提示词 p。由于去噪采样 DDIM 也是确定性的，一旦初始随机噪声 z_t 和给定提示词 p，每个步骤 t 的预测噪声 $\epsilon_\theta(z_t,\ t)$ 和最终去噪后的潜在图像 z_0 均已确定。

交叉注意力图 已知，交叉注意力图 $M(i)$ 的每个条目都对应于提示词 p [5, 7] 中的某个单词 $p(i)$。此外，生成图像中物体的布局和形状与其注意力图高度相关。简单来说，交叉注意力层计算查询 Q（即展平的输入图像 z_t 的嵌入）与键 K（即提示词 p 中每个单词 $p(i)$ 的嵌入）之间的相似度。因此，针对特定单词 $p(i)$ 的注意力图 $M(i)$ 中的值，表示输入图像中相应区域的特征与单词 $p(i)$ 的特征之间的相似度。V 包含每个单词 $p(i)$ 的丰富语义特征。根据注意力图，交叉注意力层会反复将 $p(i)$ 的特征添加到已经与单词 $p(i)$ 足够相似的区域。因此，与特定单词 $p(i)$ 相关内容的生成最终会发生在其注意力图 $M(i)$ 中具有强信号的位置。

从随机初始图像进行图像合成在扩散模型中，初始图像通常从高斯分布中采样。图像生成过程包括在潜在空间中逐渐将随机起点（初始图像）移向真实图像分布。扩散模型会去除噪声，并利用交叉注意力层计算初始图像中的每个像素与提示词中提及的每个对象之间的相似度。然后，将每个对象的特征分配到已与该对象具有相似性的区域。当初始噪声图像中的某个区域已经包含特征

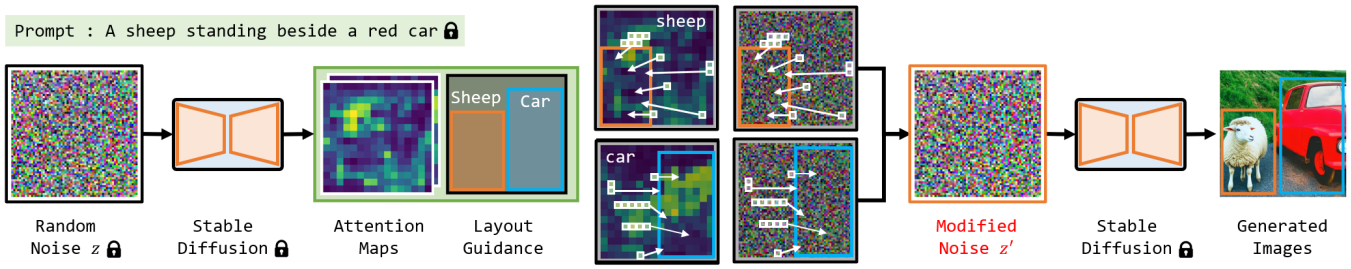

图 5：像素块交换实验的概念。我们使用初始噪声图像的注意力图来表示初始生成倾向。随后，我们将倾向于生成特定内容的像素块移动到指定区域，并像往常一样使用修改后的噪声图像进行去噪。

对于提示词中提及的特定对象，交叉注意力层会添加更多类别特征，使该区域朝着更接近该类别的方向发展，从而在该区域生成相应的对象。这种机制构成了初始图像的生成倾向。

生成失败分析根据我们的分析，导致图像生成失败的一个关键因素是初始噪声图像的生成倾向与用户提供的提示词不一致。例如，如果用户希望生成两个彼此靠近的物体，但随机生成的初始图像中，倾向于生成这两个物体的区域相距较远，模型就无法基于这样的初始图像生成所需内容。此外，一些方法[2, 17]旨在通过在注意力图中添加掩码，提高在指定区域生成指定物体的概率，从而实现布局到图像生成。然而，如果初始图像中用户指定的区域生成相关内容的倾向较低，强行在这些区域添加掩码通常会导致失败。

3.3 生成图像重绘

本节研究初始图像的部分变化对扩散模型中图像生成的影响，该方法可应用于重绘任务。

初始噪声图像在多大程度上能够决定生成图像的内容？正如上一章所讨论的，初始化图像的生成倾向与提示词之间的不一致，是导致生成失败的关键因素。因此，为避免图像生成失败，一个直观的想法是移除与提示词相冲突的生成倾向部分。首先，我们使用提示词 p 和一个随机采样的初始化图像 z 生成图像。随后，我们确定生成图像中与 p 的描述不匹配或不合理的区域。对于初始化图像 z 中与失败区域相对应的区域，我们进行重新随机化，同时保留 z 其他区域的值。我们使用部分重新随机化的初始噪声图像 $z\prime$，在相同的提示词 p 下再次执行生成，并观察生成图像。如图 3 所示，该实验简单直观，却展现出显著效果。使用部分重新随机化的初始图像 $z\prime$，能够在几乎保持图像其他区域不变的同时，自发地仅重新生成指定区域。这一结果表明，初始化图像中各区域的生成倾向具有相对独立性，对初始化图像的一小部分进行修改并不会影响其余部分的生成倾向。该实验表明，通过编辑初始化图像，我们有望以具有方向性且可控的方式影响图像生成结果。

生成图像重绘任务的应用扩散模型应用中的一个典型场景是，用户为了获得符合特定提示词的满意图像，必须进行数十次甚至数百次生成。每次生成都会使用从高斯分布中重新采样得到的初始图像，因此每次生成都是独立随机的。然而，有时用户对生成图像的大部分内容都很满意，只有其中一小部分不尽如人意。在这种情况下，用户只能从头开始重新生成图像。我们的发现可以自然地应用于这一场景。当用户对生成图像的大部分内容（例如成功生成的物体、满意的布局、色调和风格等）感到满意时，只需选出不满意的部分，在初始图像中对相应的像素值重新采样，然后使用重新采样后的图像进行生成即可。用户可以重复这一过程，直到获得满意的结果，而无需担心其他满意区域出现意外变化。这种简单的方法可以帮助用户更快地获得满意的图像。图 4 展示了一些重绘样例。

3.4 布局到图像合成

在本节中，我们将研究结果应用于布局到图像合成任务，并进一步探究初始随机图像中像素块的生成倾向这一特性。先前的实验（第 3.1 节）表明，随机初始图像中的区域倾向于生成特定内容。进一步的实验（第 3.3 节）显示，改变初始噪声图像的像素值，可以相应地改变生成内容在确切位置上的表现。这表明，初始噪声图像中各区域的生成倾向主要由其中像素点的值决定，而非其位置。这一发现可应用于布局到图像合成任务：在该任务中，提示词中提及的物体需要在用户指定的区域内生成。

问题设定 在布局到图像生成任务中，不仅给定了提示词 p，还提供了对象级引导 $\{B,\ C\}$，其中 B是指定区域列表，C是类别列表。$\{b,\ c\} \in \{B,\ C\}$ 表示应在区域 b中生成对象 c。每个 $c \in C$都应在提示词 p中提及。由于我们发现，初始噪声图像中的区域倾向于生成特定类别，因此一种直接且直观的方法是

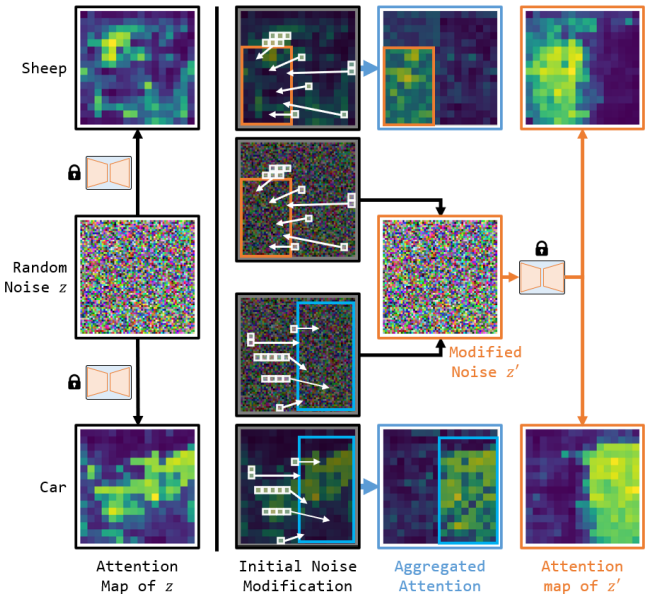

图 6：交换后的注意力图，以及通过修改后的初始噪声图像 z ，计算得到的注意力图。聚合后的注意力图显示了原始注意力图交换后的结果。 z ′ 是通过将注意力值较高的像素块交换到指定区域而得到的。 z ′ 的注意力图显示，高注意力像素块聚集的区域具有较高的生成倾向，会生成相应的内容。

将所有倾向于生成类别 c的像素块移动到指定区域 b。为此，我们需要了解初始噪声图像中像素块的生成倾向。

像素块的生成倾向 如第 3.1 节所述，U-Net 中的交叉注意力图会计算当前图像中每个像素块的生成倾向。因此，我们采用去噪第一步的注意力图来表示初始噪声图像中像素块的生成倾向。具体而言，Stable Diffusion 中使用的初始噪声图像尺寸为 64×64。为尽可能保留像素块的结构和生成倾向，我们采用下采样程度最高的交叉注意力层的注意力图（其形状为16×16）来表示初始图像中每个像素块的生成倾向；也就是说，注意力图中的每个值都对应初始噪声图像中的一个 4×4 像素块。单词 c的注意力图上数值显著的像素点，表明初始图像中与该注意力像素对应的 4×4 像素块具有生成内容 c的倾向。

像素块聚合 为简单起见，下面的解释以控制仅一个对象 b 的生成位置 c 为例，其中 b 是图像上重新缩放至 16×16的相应区域坐标。如上所述，我们可以根据注意力图，获得初始噪声图像中每个像素块的生成倾向。接下来，我们将倾向于生成 c的像素块（即在 c的注意力图上得分较高的像素块）移动到指定区域 b中。值得注意的是，我们的操作是在初始图像上进行的。我们将 z 修改为 $z\prime$ 根据注意力图，并期望仅使用 $z\prime$ 进行生成即可实现布局到图像生成。具体而言，我们计算 b的区域大小 s，即区域 b中的注意力像素数量。我们将 c 的注意力图上的所有值按从高到低排序。然后，我们选取排序后的注意力值序列中排名前 s 的值。我们用 A_{min} 表示所选第 s个值的注意力值，即其中最小的一个值。我们的目标是将所有选中的注意力值移动到指定区域 b。其中一些值可能已经位于指定区域内。我们保持这些值的位置不变，并筛选出指定区域中注意力值小于 A_{min} 的坐标，如下所示：

$$Out = \{p \mid attn(p) < A_{min},\ p \in b\} \tag{2}$$

并筛除指定区域 b 之外且注意力值大于 A_{max} 的坐标，如下所示，

$$In = \{p \mid attn(p) > A_{min},\ p \notin b\} \tag{3}$$

In 和 Out 中的元素数量相同。我们随机组合 In 和 Out 中的元素，形成交换对。 In 和 Out 中的每个元素都会被使用，且仅使用一次。

$$Swap = \{\{i, o\} \mid i \in In,\ o \in Out\} \tag{4}$$

注意力图上的每个位置都对应初始噪声图像上的一个 4×4 像素块，对于每个交换对$\{i,\ o\} \in Swap$，我们交换初始噪声图像上与 i和 o 对应的像素块。我们将 4×4 的像素块作为不变单元进行交换，保持其内部结构和值不变。随后，稳定扩散会像往常一样，在修改后的初始图像 $z\prime$ 上进行最终生成。整体流程如图5所示。当指定了多个物体$\{B,\ C\}$的区域时，为确保每个像素块都被分配给唯一的物体，我们首先根据这些像素块在各类别$c \in C$的注意力图上的取值进行分类。对于每个$\{b,\ c\} \in \{B,\ C\}$，我们将归类为 c的像素块排序并移动到相应区域 b。

移动后像素块的生成倾向如图6所示，我们将初始图像 z 中倾向于生成内容 c 的像素块移入指定区域 b，并获得修改后的初始图像 $z\prime$。随后，我们通过同一稳定扩散模型的 U-Net 获取 $z\prime$ 的注意力图。我们观察到，修改后的图像 $z\prime$ 中的指定区域 b 明显倾向于生成类别 c，这可从类别 c注意力图中对应区域 b 的高强度信号看出。这一现象意味着，这些像素块的生成倾向会随像素值的移动而改变，进一步表明像素块的生成倾向主要由像素值而非其空间位置决定。

与基于掩码的方法比较一些方法 [2, 17]试图通过人为地在指定类别的注意力图上添加掩码，提高指定类别出现在指定区域的概率。这种方法和我们的方法最终都会使指定区域的注意力图中的数值升高。然而，基于掩码的方法从根本上受限于一个事实：它们完全忽略了初始图像本身的生成倾向。如果指定区域内的初始像素本身就难以生成指定类别，那么通过基于掩码的方法强行将指定类别的特征添加到指定

方法	交并比 ↑	交并比$_s$ ↑	交并比$_m$ ↑	交并比$_l$ ↑	R成功% ↑	R^s成功% ↑	R^m成功% ↑	R^l成功% ↑	FID↓
SD [25]	0.19	0.05	0.21	0.38	12.14	0.50	8.11	39.41	20.8
Paint [2]	0.24	0.08	0.27	0.46	17.56	1.16	14.05	52.34	22.7
Soft [17]	0.27	0.10	0.31	0.48	22.31	2.95	22.38	56.24	21.3
Ours	0.26	0.09	0.32	0.46	21.58	2.26	22.75	54.21	21.7
Paint [2] +Ours	0.29	0.11	0.35	0.51	23.89	2.12	24.74	61.48	23.7
Soft [17] + 我们的方法	**0.31**	**0.12**	**0.37**	**0.52**	**26.41**	**3.24**	**29.55**	**62.59**	**21.2**

表 1：布局引导的控制效果。单独使用我们的方法时，其性能可与最先进的方法相媲美；将我们的方法与这些方法结合使用，则能在所有子集上取得最佳性能。

类别在指定区域内，很可能导致生成失败或影响图像质量。另一方面，如果指定区域以外的其他区域本身就倾向于生成指定类别，那么即使在指定区域添加掩码，也无法阻止区域外物体的生成。与这些方法不同，移动初始化图像的像素块从根本上改变了初始化图像的生成倾向。即使没有施加额外干预，模型也会在后续生成过程中自发地在指定区域生成指定物体。

4 评估

为验证我们的方法在布局到图像合成任务中的有效性，并展示我们方法的优势，我们对方法进行了定量评估，并将其与当前最先进的方法进行比较。我们还通过将基于掩码的方法与我们的方法相结合，开展了提升其性能的实验。实验设置遵循 Soft-Mask [17], , 其中利用物体检测器评估逐物体控制的效率。

4.1 数据集

我们在 MSCOCO [14]数据集上进行了定量评估，该数据集包含图像描述和物体标注。我们选取了边界框类别均在其图像描述中提及的样本。由于大多数图像描述仅涉及图像中的一个或两个显著物体，因此我们随机选取了 6000 张图像，其中 3000 张包含一个物体，另外 3000 张包含两个物体，与先前工作保持一致。所有图像及其对应的标注均调整为 512_2。在去噪过程中，图像的描述被用作提示词，而物体标注则被用作布局引导。

4.2 指标

布局控制的有效性 我们使用图像描述作为提示词，并将逐物体标注作为布局引导来生成图像。随后，我们使用物体检测器 YOLOR [30], 对生成图像进行物体检测，获取生成物体的类别和检测到的边界框。检测结果与布局引导的一致性可以体现布局控制的有效性。对于每个物体，我们计算其布局引导（即数据集中的边界框标注）与检测到的边界框之间的交并比。若交并比大于 0.5，则视为控制成功，成功率表示为Rsuc。为了进一步评估对不同尺寸物体的控制效果，我们根据物体的面积及其与图像中心的距离，将所有物体划分为若干子集。具体而言，子集 s 包含面积小于 150_2的物体，

子集 l 包含所有面积大于 300_2的物体，其余物体则归入子集 m。

图像质量弗雷歇起始距离 [8] 用于衡量两组图像之间的相似度，通常用于表征生成图像的质量。我们通过生成图像和真实图像（即数据集中的图像）计算弗雷歇起始距离。弗雷歇起始距离越小，图像质量越好。

4.3 实现

我们采用默认配置的稳定扩散（SD） [25] v1.4 作为本研究的基准模型。在此设置下，模型使用 PLMS 执行 50 个去噪采样步骤。基准稳定扩散模型不接受任何布局引导，但可能会偶然在指定区域生成指定物体。在其他实验的每个图像生成过程中，模型都会接收提示词和布局引导，并生成相应的图像。我们的方法使用模型的 U-Net，根据随机采样的初始图像和提示词 z, p 计算交叉注意力图。随后，我们通过交换操作，将 z中注意力得分较高的像素块移动到指定区域，从而获得经过修改的初始噪声图像 z_r。然后，我们使用 $\{z_r, p\}$ 执行常规去噪。我们将所提出的方法与 Paint [2] 和 Soft [17] 进行比较。这两种方法都会在用户指定的区域添加掩码，以增强相应区域中的注意力值。Soft [17] 还会抑制区域外的注意力。我们使用相同的 $\{z, p\}$ 进行去噪，并在每个去噪步骤中，将布局掩码添加到每个 U-Net 层的交叉注意力图中。此外，我们还报告了将本方法与基于掩码的方法相结合时的性能。在这种情况下，我们使用 $\{z_r, p\}$ 生成图像，并在去噪过程中使用掩码。

4.4 实验结果与分析

评估结果如表 1 所示。我们的方法在所有子集上的表现均显著优于基线方法 [25] 和 Paint [2]，且无需对模型的去噪过程施加任何干扰，便能达到与 Soft [17], 相当的性能。此外，我们的方法与基于掩码的方法作用于生成过程的不同阶段。若将去噪过程视为在潜在空间中将一个随机起点移动到目标点，那么我们的方法可以看作是为去噪过程选择了一个更好的起点。与此同时，基于掩码的方法可以视为在移动过程中提供一个朝向固定方向的额外引导。因此，这两种方法并不冲突，而是可以自然地结合使用。直观来看，与我们的方法结合能显著提升两种基于掩码的方法的性能，而我们的方法与 Soft [17] 的组合在所有子集上均取得了最佳性能。一些高质量样例

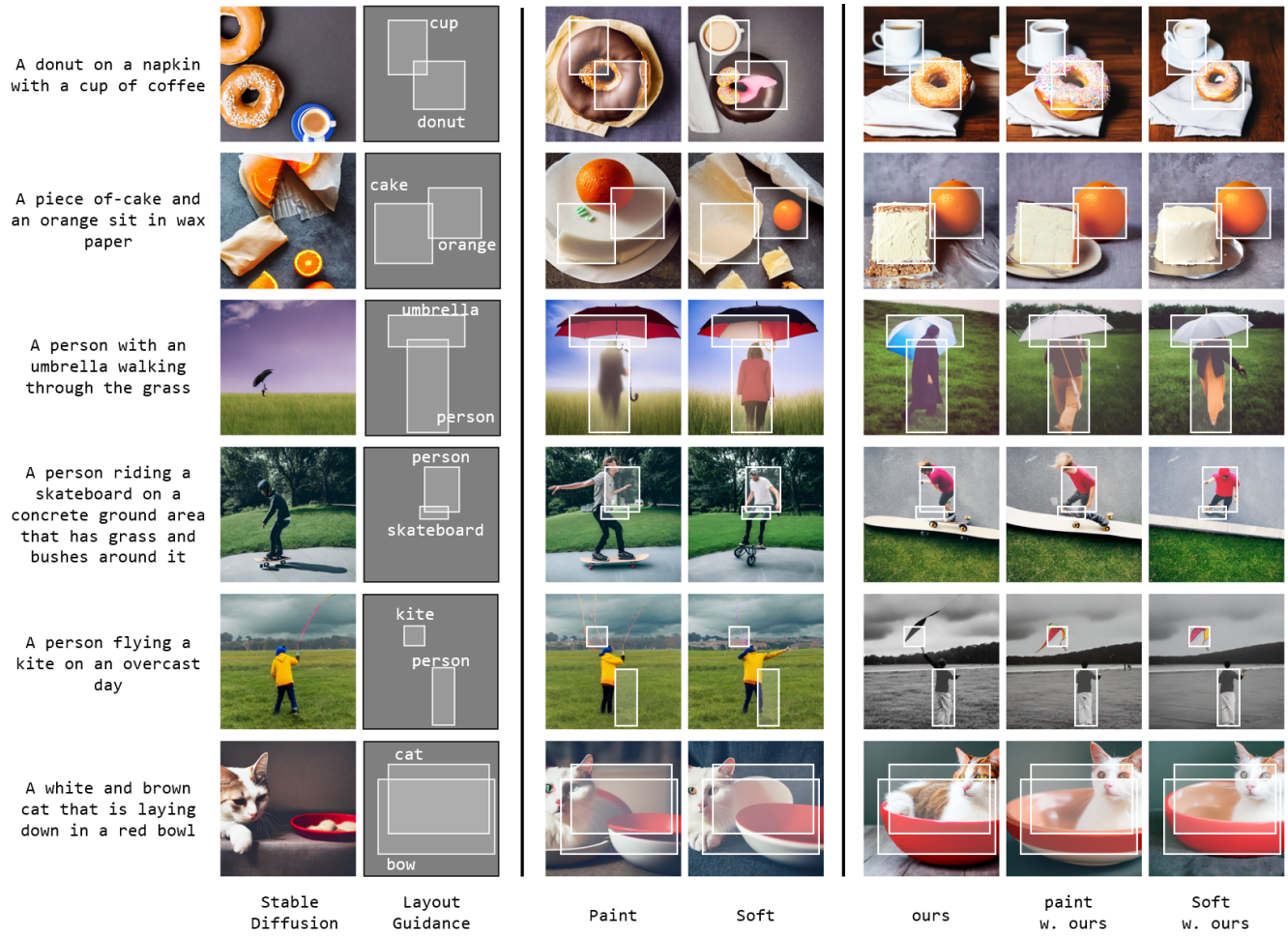

图 7：我们在 MS COCO 上实验得到的高质量样例。每行中的图像均由同一张初始噪声图像生成。在大多数情况下，即使应用了不同的掩码，未使用我们方法的图像仍有显著的共同点（结构、背景、风格）。应用我们的方法后，像素块的部分位置会互换，初始图像也会发生显著变化。因此，生成图像也会随之改变。

如图 7 所示。我们的方法的有效性表明，初始随机图像的生成倾向主要取决于像素值，而非空间位置，并且生成倾向会随着初始图像中像素值的移动而变化。我们方法的一个局限性在于，当引导边界框较小时，我们的方法只能移动初始图像中的少量像素，有时无法对生成结果产生显著影响。因此，与较大的物体相比，我们的方法对较小物体的效果相对较弱。

5 结论与未来工作

本文提出了一种通过修改随机采样的初始噪声图像来实现扩散模型图像生成细粒度控制的新方向。尽管初始噪声图像传统上被认为是不包含任何有意义信息的随机噪声信号，但我们的研究发现，当其作为去噪的起点时，这类图像具有在特定位置生成特定内容的内在倾向。我们研究了促成这种倾向的潜在因素，并尝试调整初始图像的生成倾向，使其影响最终输出。实验结果表明，改变初始噪声图像的某个特定区域确实会影响生成图像中的相应内容，我们将这一特性应用于重绘任务。此外，我们发现，这种生成倾向主要取决于像素值，而非空间位置；随着初始图像中像素值的移动，该倾向也会发生变化。我们将这一发现应用于布局到图像生成任务，并取得了最先进的性能。基于本文的研究结果，我们发现了若干令人振奋的潜在研究方向。具体而言，我们的方法凸显了初始图像在图像生成过程中的重要性。因此，围绕初始图像优化，或基于优化后的初始图像加速去噪的研究，具有巨大的发展潜力。

致谢

本研究部分获得了 JST Mirai-Program JP- MJMI21H1 的资助。

参考文献

[1] Omri Avrahami、Thomas Hayes、Oran Gafni、Sonal Gupta、Yaniv Taigman、Devi Parikh、Dani Lischinski、Ohad Fried 和 Xi Yin。2022 年。SpaText：用于可控图像生成的时空文本表征。*arXiv* 预印本 *arXiv:2211.14305*（2022 年）。[2] Yogesh Balaji、Seungjun Nah、Xun Huang、Arash Vahdat、Jiaming Song、Karsten Kreis、Miika Aittala、Timo Aila、Samuli Laine、Bryan Catanzaro 等。2022 年。ediffi：采用专家去噪器集成的文本到图像扩散模型。*arXiv* 预印本 *arXiv:2211.01324*（2022 年）。[3] Prafulla Dhariwal 和 Alexander Nichol。2021 年。扩散模型在图像合成方面优于生成对抗网络。*NeurIPS* 34（2021 年），8780–8794。[4] Ming Ding、Zhuoyi Yang、Wenyi Hong、Wendi Zheng、Chang Zhou、Da Yin、Junyang Lin、Xu Zou、Zhou Shao、Hongxia Yang 等。2021 年。Cogview：通过 Transformer 掌握文本到图像生成。*NeurIPS* 34（2021 年），19822–19835。[5] Weixi Feng、Xuehai He、Tsu-Jui Fu、Varun Jampani、Arjun Akula、Pradyumna Narayana、Sugato Basu、Xin Eric Wang 和 William Yang Wang。2022 年。无需训练的结构化扩散引导，用于组合式文本到图像合成。*arXiv* 预印本 *arXiv:2212.05032*（2022 年）。[6] Oran Gafni、Adam Polyak、Oron Ashual、Shelly Sheynin、Devi Parikh 和 Yaniv Taigman。2022 年。Make-a-scene：利用人类先验的基于场景的文本到图像生成。载于*ECCV*。施普林格，89–106。[7] Amir Hertz、Ron Mokady、Jay Tenenbaum、Kfir Aberman、Yael Pritch 和 Daniel Cohen-Or。2022 年。通过交叉注意力控制实现提示词到提示词图像编辑。*arXiv* 预印本 *arXiv:2208.01626*（2022 年）。[8] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。2017 年。采用双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。*NeurIPS* 30（2017 年）。[9] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020 年。去噪扩散概率模型。*NeurIPS* 33（2020 年），6840–6851。[10] Jonathan Ho 和 Tim Salimans。[无日期]。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。[11] Lianghua Huang、Di Chen、Yu Liu、Yujun Shen、Deli Zhao 和 Jingren Zhou。2023 年。Composer：通过可组合条件实现创造性、可控的图像合成。*arXiv* 预印本 *arXiv:2302.09778*（2023 年）。[12] Bahjat Kawar、Shiran Zada、Oran Lang、Omer Tov、Huiwen Chang、Tali Dekel、Inbar Mosseri 和 Michal Irani。2022 年。Imagic：使用扩散模型进行基于文本的真实图像编辑。*arXiv* 预印本 *arXiv:2210.09276*（2022 年）。[13] Gwanghyun Kim、Taesung Kwon 和 Jong Chul Ye。2022 年。Diffusionclip：用于鲁棒图像操控的文本引导扩散模型。载于*CVPR*。2426–2435。[14] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014 年。Microsoft COCO：上下文中的常见物体。载于*ECCV*。施普林格，740–755。[15] Luping Liu、Yi Ren、Zhijie Lin 和 Zhou Zhao。2022 年。流形上扩散模型的伪数值方法。载于*ICLR*。[16] Nan Liu、Shuang Li、Yilun Du、Antonio Torralba 和 Joshua B Tenenbaum。2022 年。利用可组合扩散模型进行组合式视觉生成。载于*ECCV*。施普普林格，423–439。[17] Jiafeng Mao 和 Xueting Wang。2023年。无需训练的位置感知文本到图像合成。*arXiv* 预印本 *arXiv:2304.13427*（2023）。[18] Chong Mou、Xintao Wang、Liangbin Xie、Jian Zhang、Zhongang Qi、Ying Shan 和 Xiaohu Qie。2023年。T2I-Adapter：通过学习适配器挖掘文本到图像扩散模型中更多可控能力。*arXiv* 电子论文（2023），arXiv–2302。[19] Alexander Quinn Nichol 和 Prafulla Dhariwal。2021年。改进的去噪扩散概率模型。载于 *ICML*。PMLR，8162–8171。[20] Alexander Quinn Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob Mcgrew、Ilya Sutskever 和 Mark Chen。2022年。GLIDE：利用文本引导的扩散模型实现逼真图像生成与编辑。载于 *ICML*。PMLR，16784–16804。[21] Dong Huk Park、Samaneh Azadi、Xihui Liu、Trevor Darrell 和 Anna Rohrbach。2021年。组合式文本到图像合成基准。载于 *NeurIPS* 数据集与基准赛道（第一轮）。[22] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。从自然语言监督中学习可迁移的视觉模型。载于 *ICML*。PMLR，8748–8763。[23] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。利用 CLIP 潜变量进行层次化文本条件图像生成。*arXiv* 预印本 *arXiv:2204.06125*（2022）。[24] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。2021年。零样本文本到图像生成。载于 *ICML*。PMLR，8821–8831。[25] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022年。利用潜扩散模型进行高分辨率图像合成。载于 *CVPR*。10684–10695。[26] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。2015年。U-Net：用于生物医学图像分割的卷积网络。载于 *MICCAI*。施普林格，234–241。[27] Jiaming Song、Chenlin Meng 和 Stefano Ermon。2021年。去噪扩散隐式模型。载于 *ICLR*。[28] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力机制就是你所需要的一切。*NeurIPS* 30（2017）。[29] Andrey Voynov、Kfir Aberman 和 Daniel Cohen-Or。2022年。草图引导的文本到图像扩散模型。*arXiv* 预印本 *arXiv:2211.13752*（2022）。[30] Chien-Yao Wang、I-Hau Yeh 和 Hong-Yuan Mark Liao。2021年。你只需学习一种表示：用于多任务的统一网络。*arXiv* 预印本 *arXiv:2105.04206*（2021）。[31] Tengfei Wang、Ting Zhang、Bo Zhang、Hao Ouyang、Dong Chen、Qifeng Chen 和 Fang Wen。2022年。预训练就是图像到图像转换所需的一切。*arXiv* 预印本 *arXiv:2205.12952*（2022）。[32] Lvmin Zhang 和 Maneesh Agrawala。2023年。为文本到图像扩散模型添加条件控制。*arXiv* 预印本 *arXiv:2302.05543*（2023）。