Computer Science and Application 计算机科学与应用, 2025, 15(4), 443-452
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154116

Hans汉斯

基于布局控制的文本到图像扩散模型研究进展

齐时达

温州大学计算机与人工智能学院，浙江 温州

收稿日期：2025年3月25日；录用日期：2025年4月23日；发布日期：2025年4月30日

摘 要

随着计算机视觉和生成模型的迅猛发展，布局到图像生成(Layout-to-Image Generation)已成为一个重要的研究方向。该任务通过提供物体的空间布局信息，如边界框位置和类别标签，生成符合该布局要求的真实图像。近年来，扩散模型作为一种新兴的生成技术，凭借其在图像生成中的独特优势，逐渐成为布局到图像生成的主流方法之一。与生成对抗网络(GAN)相比，扩散模型在图像质量、稳定性和多样性方面表现出更好的性能。本文综述了近年来扩散模型在布局到图像生成中的研究进展，详细介绍了扩散模型的基本原理，并将现有的研究成果归纳为三类：1) 专用扩散模型方法；2) 基于预训练扩散模型的适配方法；3) 推理阶段的组合控制方法。本文还分析了不同布局生成方法的优缺点，并对未来可能的研究方向进行了展望。

关键词

扩散模型，布局控制，生成对抗网络，图像生成，预训练

Research Progress on Text-to-Image Diffusion Models Based on Layout Control

Shida Qi

College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Mar. 25th, 2025; accepted: Apr. 23rd, 2025; published: Apr. 30th, 2025

Abstract

With the rapid development of computer vision and generative models, Layout-to-Image Generation has become an important research direction. This task involves generating realistic images that conform to the given spatial layout of objects, such as bounding box positions and class labels. In recent years, diffusion models, as an emerging generative technique, have gradually become one of the main methods for Layout-to-Image Generation due to their unique advantages in image

文章引用: 齐时达. 基于布局控制的文本到图像扩散模型研究进展[J]. 计算机科学与应用, 2025, 15(4): 443-452.
DOI: 10.12677/csa.2025.154116

generation. Compared to Generative Adversarial Networks (GANs), diffusion models perform better in terms of image quality, stability, and diversity. This paper reviews the recent advancements of diffusion models in Layout-to-Image Generation, provides a detailed introduction to the fundamental principles of diffusion models, and categorizes the existing research into three types: 1) Dedicated diffusion model methods; 2) Adaptation methods based on pre-trained diffusion models; 3) Combination control methods during the inference stage. The paper also analyzes the advantages and disadvantages of different layout generation methods and discusses potential future research directions.

Keywords

Diffusion Models, Layout Control, Generative Adversarial Networks, Image Generation, Pre-Training

1. 引言

在计算机视觉生成领域，从布局(layout)出发生成图像是一个重要而具有挑战性的任务。所谓布局到图像(Layout-to-Image)生成，指的是给定图像中各物体的空间布局(例如，每个物体的边界框位置及类别标签)，让模型合成符合该布局要求的真实图像。这一任务允许用户在生成图像时精确控制物体的种类及其在图中的位置，因此在内容创作、交互式设计等应用中具有巨大潜力。例如，设计师可以先用粗略的矩形布局指定场景中主要物体的位置，然后让生成模型填充出逼真的图像[1]。相比仅根据文本描述生成图像，基于布局的控制可以避免文字描述中的歧义，更直接地指定空间结构，从而在多物体复杂场景下提供更高的控制精度[2]。

然而，从布局生成图像也面临多方面挑战。首先，模型需要确保生成图像的感知质量足够高，即图像清晰逼真，这对复杂场景尤为困难[3]。其次，模型必须严格遵循布局约束，确保每个物体出现在指定的位置和大小，并且不同物体之间不相互重叠冲突。布局中的物体往往存在类别多样性和尺度变化，模型既要保持布局控制，又要生成各异且合理的物体外观，这需要在多样性与一致性之间取得平衡。传统的生成对抗网络(GAN) [4]-[8]方法在此任务上取得了一定进展，但由于训练不稳定[9]和模式坍塌[10]等问题，难以兼顾图像质量和布局精确度。近年来，新兴的扩散模型(Diffusion Models, DMs) [11]为多物体图像生成带来了突破性的进步，大量研究将扩散模型引入布局到图像生成领域，取得了远超以往 GAN 方法的性能[12]。本文将对近年来扩散模型在布局到图像(Layout-to-Image)生成领域的研究进展进行全面综述。本文将首先介绍扩散模型和布局到图像任务的背景知识，然后系统归纳该领域的方法，按照其策略分为三类进行梳理，包括专用扩散模型方法、基于预训练扩散模型的适配方法、以及推理阶段的组合与控制方法。接着，我们将比较不同方法的性能与优缺点。最后，我们讨论当前存在的挑战并对其未来发展进行了展望。

2. 相关技术背景

2.1. 布局到图像任务的早期方法

布局到图像生成要求模型从抽象布局信息合成完整图像，其输入通常表示为一组带有类别标签的边界框(bounding boxes)或像素级语义分割掩码。这一任务代表性工作是“Layout2Im”模型[13]，Layout2Im

将布局定义为若干目标对象的类别和位置，采用了 VAE [14] + GAN 架构：通过 VAE 编码物体外观、LSTM 融合布局顺序、以及图像解码器重构图像，并使用全局和对象级对抗损失来提高生成效果。该模型证明了直接从布局生成图像的可行性，比之前将布局仅作为中间表示的方法生成结果更加准确且多样。随后，多种 GAN 框架被应用于该任务：例如 LostGAN 系列引入了可重构布局的概念，提出了 LostGAN-v1 [15]并改进为 LostGAN-v2 [16]。LostGAN 通过在生成过程中灵活调整布局和风格编码，实现对各对象外观和位置的更好控制。OC-GAN (Object-Centric GAN) [17]强调对场景中对象及其关系的理解，采用场景图相似性模块和改进的条件机制来缓解 GAN 生成中的常见错误，如虚假物体或重叠物体。另一项工作 PLGAN (Panoptic Layout GAN) [7]将“东西(stuff)”和“实例(instance)”分割概念引入布局生成，通过双分支网络分别处理非物体背景和具体物体，并设计实例/背景感知的归一化融合为全景布局，从而提升多物体场景的真实性。总体而言，这一时期的 GAN 方法在布局遵循性上有所进步，但仍存在生成质量欠佳、模式坍塌[10]和训练不稳定[9]等局限。例如，GAN 模型在复杂布局下容易生成失真或合并的物体[2]。随着扩散模型的兴起，研究者开始尝试用扩散模型替代 GAN 来解决布局到图像任务中的难点。

2.2. 扩散模型原理简述

扩散模型是一类基于概率扩散过程的生成模型。它包含一个正向扩散过程，将训练数据逐步添加噪声直至接近各向同性高斯分布，以及一个学习得到的逆向去噪过程，用于从纯噪声逐步还原数据分布。以图像生成为例，正向过程定义为：在时间步 t 中，从真实图像 x_0 出发逐步生成带噪图像 x_t，常用线性调度的噪声增量 β_t 控制每步扰动强度。公式表示为：

$$q\left(x_t \middle| x_{t-1}\right)=\mathcal{N}\left(x_t;\ \sqrt{1-\beta_t}\,x_{t-1},\ \beta_t\mathbf{I}\right),\quad t=1,\cdots,T \tag{1}$$

其中 T 是总扩散步数。当 $t=T$ 足够大时，x_T 近似服从各向同性高斯噪声。扩散模型的学习目标是训练一个参数化模型 p_θ 去逼近逆过程 $p\left(x_{t-1}\middle|x_t\right)$。Ho 等人提出了等价的简化训练目标，即让模型 $\epsilon_\theta\left(x_t,t,c\right)$ 去预测加入噪声时的噪声分量 ϵ。对于有条件生成(条件 c 可以是标签、文本或布局等)的情形，常用的损失函数形式为：

$$L_{\text{simple}}=E_{x_0,\epsilon,t}\left[\left\|\epsilon-\epsilon_\theta\left(x_t,t\middle|c\right)\right\|^2\right] \tag{2}$$

其中 $\epsilon\sim\mathcal{N}\left(0,I\right)$，$x_t=\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}$，($\bar{\alpha}_t$ 为噪声超参数的累积)表示由真实图像加噪得到的训练样本。通过最小化上述均方误差，UNet 结构的去噪网络 ϵ_θ 学会了在给定条件 c 和含噪图像 x_t 时，逐步去除噪声复原图像。由于扩散模型采用对数似然训练，不存在 GAN 的模式坍塌问题，并且在捕获数据分布多样性和语义一致性方面表现更佳。

在采样时，模型从高斯噪声 $x_T\sim\mathcal{N}\left(0,\boldsymbol{I}\right)$ 出发，通过迭代去噪将高斯噪声逐步转化为目标图像。在每一步 $t\to t-1$，根据当前带噪图像 x_t 预测去噪输出 $\widehat{x_{t-1}}$。常用的 PNDM [18]、DDIM [19]等加速采样方法可以在保持图像质量的同时显著减少步数，例如 LayoutDiffusion [2]利用高阶 ODE 求解器将采样步数减至 25 步仍能取得优异效果。此外，无分类器引导(classifier-free guidance) [20]是扩散模型常用的条件采样增强技术。其做法是同时训练有条件和无条件两个网络分支，然后在采样时将两者输出线性组合，从而放大条件信息的影响力：

$$\hat{\epsilon}_\theta\left(x_t\middle|c\right)=w\cdot\epsilon_\theta\left(x_t\mid c\right)+\left(1-w\right)\cdot\epsilon_\theta\left(x_t\middle|\varnothing\right) \tag{3}$$

其中 $\epsilon_\theta\left(x_t\varnothing\right)$ 表示忽略条件时预测的噪声，$w>1$ 为引导系数。当 w 增大时，生成结果对条件 c 的符合度提高，但也可能牺牲一定写实度。该技术已成功用于条件扩散模型的提升，如 Imagen、Stable Diffusion 等皆采用了 $w=7.5$ 左右的引导配置来获得更符合文本/标签条件的图像。总的来说，扩散模型为布局到图像生成带来了稳健的训练和高保真多样的样本质量。本章着重介绍了布局到图像生成的早期方法和扩散模

型的基本原理，接下来本文将深入探讨近年基于扩散模型的布局到图像方法，并分析其核心思想。

3. 基于扩散模型的布局到图像生成方法

近年来的研究可大致分为以下几类：一类是专门为布局生成设计并从零训练的扩散模型，强调针对多物体布局的结构建模；第二类是利用预训练扩散模型并进行适配微调的方法，通过引入额外模块将大规模预训练的扩散模型应用于布局条件生成；第三类是推理阶段的组合与控制方法，不需要大量重新训练，通过巧妙的推理算法实现布局控制，包括注意力引导、迭代生成和噪声拼接等。下面我们按类别介绍主要方法，每类中按时间顺序或方法关联进行组织，并重点比较它们的关系和差异。

3.1. 专用扩散模型方法

LayoutDiffusion [2]是这一方向的代表性工作之一。该模型由 Zheng 等人提出，旨在以扩散模型实现对复杂多物体布局的高保真图像生成和精细控制。LayoutDiffusion 采用了条件 DDPM 框架，从随机噪声迭代生成图像，并针对布局条件设计了两大关键组件：布局融合模块(LFM)和物体感知交叉注意力(Object-aware Cross-Attention, OaCA)。整体思路是将输入布局和图像表示转化为统一空间下的信息融合，增强多物体之间的关系建模。具体而言，对于输入的 n 个物体布局 $o_1,\cdots,o_n$，每个物体包含边界框 $b_i=(x_i, y_i, w_i, h_i)$ 位置和类别 c_i。LayoutDiffusion 首先通过可学习映射将每个边界框坐标编码为向量 $B_{L,i}$，将类别标签编码为向量 $C_{L,i}$，并将二者相加得到物体的初始嵌入 $O_i=B_{L,i}+C_{L,i}$。公式表示整个布局序列嵌入为：

$$L=\{O_1,O_2,\cdots,O_n\},\ O_i=W_B b_i+W_C c_i \tag{4}$$

其中 W_B, W_C 分别为边界框和类别的嵌入矩阵。通过这样的L向量序列表示，模型获得了同时包含物体类别(内容)和位置(坐标)信息的布局表示。接下来，布局融合模块 LFM 对 L 应用多层 Transformer 自注意力，使布局序列内部的信息充分交互，得到融合后的布局表示 L'。这一步类似于构建一个“布局图”的过程，让模型理解布局中各对象之间的相对关系(如遮挡、邻近)。然后，在实际的图像生成 UNet 中，作者将扩散模型中中间层的特征图划分为若干图像 patch，并视每个 patch 为特殊的“对象”。这些图像 patch 也被赋予与其空间位置对应的嵌入表示，与布局嵌入映射到统一坐标系下。通过这样设计，布局和图像特征可以在相同空间中进行融合。物体感知交叉注意力 OaCA 模块就是在这一阶段作用：扩散 UNet 某层以图像 patch 特征为查询(query)、布局融合表示 L' 为键(key)和值(value)，执行 Cross-Attention 计算。这一机制使生成网络在每个位置关注对应的布局对象信息，从而精细地将特定物体特征注入图像。特别地，由于 key、value 中包含物体类别和位置嵌入，注意力能够引导 UNet 在正确的位置生成具有正确类别的内容。OaCA 被称为“物体感知”，因为相比常规跨模态注意力，它专门针对布局中每个对象进行局部强化。此外，LayoutDiffusion 在训练中采用分类器自由指导来增强布局条件效果，并对采样过程进行优化，在 25 步采样内即可明显优于 GAN 基准。可以说，LayoutDiffusion 证明了专门设计的扩散模型在多物体布局生成上的巨大潜力。其优点是从架构上针对布局融合进行了优化，生成结果在复杂场景下依然质量高且布局精确。但其局限在于仍需要在特定数据集上从头训练大规模模型，模型认知范围受限于训练语料(例如 COCO 的 80 类物体)。对于训练集中未出现的新类别或新场景，模型难以胜任，这引出了下一节所述的利用预训练扩散模型的方法。

3.2. 基于预训练扩散模型的适配方法

与从零训练模型不同，另一类思路是利用预训练的扩散模型(通常是在大规模数据上训练的文本到图像扩散模型或图像扩散模型)，通过增加少量参数或调整网络，使其能够接受布局条件输入。这类方法的动机是：大规模预训练模型已经掌握了丰富的视觉概念和生成能力，如果能在不遗忘的前提下接入布局

控制，将同时实现开放世界概念和精细布局控制。这一思想借鉴了识别领域的“迁移学习”范式，即从基础模型出发适应下游任务。

LayoutDiffuse [1]由 Cheng 等人提出，旨在高效地将预训练扩散模型适配为布局到图像生成器。作者提出，直接将布局信息拼接成类似文本序列输入预训练模型会遇到分布不匹配的问题。为此，LayoutDiffuse 采用在模型内部加入轻量适配器(adapter)的策略。具体做法是在预训练扩散 UNet 的中间各层插入小的残差模块，这些模块接收布局作为额外输入，并学习调整原模型的中间特征。为了高效融合布局，LayoutDiffuse 引入了布局注意力(layout attention)和任务自适应提示(task-aware prompt)两个机制。布局注意力类似前述 OaCA 思想，将自注意力限制在每个实例内部，强化模型对同一物体区域像素的关联建模，同时引入全局类别嵌入捕捉不同物体间的关联。任务自适应提示则是在模型输入中加入可学习向量，作为引导预训练模型切换到“布局生成模式”的开关。这些提示向量在微调中被调整，使模型逐步适应仅有布局而无文本时也能正常生成。通过这两部分的适配模块，LayoutDiffuse 只需要冻结原有扩散模型权重，训练少量参数，就实现了模型的快速收敛和对布局的准确条件控制。其优势在于训练高效、数据高效，充分利用了已有模型的知识。相较从头训练，适配方法避免了重复训练庞大网络的代价。然而，该方法目前仍局限于已有模型的知识范围，对于非常规布局可能存在一定偏差，因为基础扩散模型在预训练时并未见过大量极端布局配置(例如物体密集堆叠等)。

另一项具有影响力的工作是微软提出的 GLIGEN (Grounded-Language-to-Image Generation) [21]。GLIGEN 关注开放集场景下的布局控制，即不局限于固定类别标签，而允许用自然语言描述任意物体并指定其位置。GLIGEN 基于预训练的 Stable Diffusion 模型，采取的方法是添加门控的 Transformer 层以融合新输入(如边界框和文字)。具体而言，GLIGEN 在 Stable Diffusion 原有 UNet 的多层交叉注意力模块中插入新的注意力层，这些层以区域坐标等作为附加输入，通过可训练的门控单元与原模型输出融合。训练时，原 Stable Diffusion 权重保持冻结，仅训练新加的门控 Transformer 层。这种设计确保了原模型对大规模图像–文本知识的保留，避免了“灾难性遗忘”问题。GLIGEN 的输入包括文本描述和对应边界框坐标，模型能够在采样早期融合文本和布局信息，在采样后期逐渐仅依靠原模型保证图像质量。这种两阶段的采样策略(前半段使用全模型含门控层，后半段仅用冻结的原模型层)巧妙地实现了控制与质量的兼顾。在不额外监督的情况下，预训练扩散模型经过适配可以零样本地完成布局控制任务，并且效果显著超过以往有监督训练的专用模型。GLIGEN 的优势是能处理开放物体集和更加自由的描述，但相应地，它需要文本描述和布局同时作为条件。这意味着在使用 GLIGEN 时，用户需要为每个定位框提供文本标签。这在开放世界场景是合理的(因为类别可能不限于封闭集合)，但在封闭场景下则增加了额外的信息需求。

ControlNet [22]也是需要提及的相关进展。ControlNet 并非专为布局生成而提出，但作为一种通用条件控制扩散模型的方法，对比很有意义。ControlNet 核心思想是在不改动原有扩散模型权重的前提下，复制一份 UNet 结构并将其作为“条件分支”进行训练。该条件分支接收各种外部条件(如边缘图、姿态骨架、分割图等)，通过在各 UNet 层注入与原模型相同尺寸的特征来影响采样过程。训练时冻结原模型，只训练条件分支，使其输出作为条件引导融入到原扩散网络，从而实现对最终生成的额外可控。对于布局到图像任务，ControlNet 可以接受语义分割掩码作为控制(相当于一种像素级布局)。ControlNet 的优点在于通用性强、可扩展：可以针对不同条件训练不同的分支而不影响主模型，并且一个预训练模型可以配备多个 ControlNet 以组合多种条件。它的效果同样依赖预训练模型的知识，并通过条件网络实现精细控制。

除了上述工作外，InstanceDiffusion [23]通过学习可调节的 UniFusion 模块，通过动态分割掩码增强模型对重叠物体的建模能力。MS-Diffusion [24]设计了多尺度噪声预测网络，在低分辨率阶段学习全局布局约束，高分辨率阶段细化局部细节，显著提升了复杂场景的生成一致性。它们共同的优势是借助大规

模预训练模型的知识，实现了比从头训练方法更强的生成能力和概念泛化能力。不足之处在于仍需一定的微调开销，并且如果布局要求超出预训练模型经验(如非常规的物体排列)，可能出现一定偏差或需要特殊处理(如 GLIGEN 通过文字提醒模型注意新场景)。

3.3. 推理与组合控制方法

除了修改模型结构和训练方式外，还有一类研究聚焦于推理阶段的算法设计，以实现对布局的控制和优化，而不需要大规模模型改动。这类方法通常利用预训练扩散模型，通过多次推理、迭代调整或直接操纵扩散过程来满足布局约束。其好处是无需(或只需很少)额外训练，灵活性高，特别适合处理训练分布之外的布局情形。典型的方法包括迭代生成/局部重绘、注意力引导以及噪声合成等。

Iterative Inpainting (迭代填充) [3]思路是由 Cho 等人在 2023 年提出的，该方法针对模型在分布外布局(OOD)上性能下降的问题，提出逐步生成图像：先生成背景，再一个一个地将布局中的物体叠加到图像中，每次生成一个物体区域并进行图像局部填充。具体而言，对于给定布局，IterInpaint 按照某种顺序选择一个物体，将除该物体区域外的其它区域掩膜，然后利用扩散模型的图像修复(inpainting)能力在该空白区域生成该物体。生成完一个物体后，更新图像，再继续下一个物体，直至所有物体都被绘制出来。这样，模型每次只需关注当前一个物体及其局部环境，降低了单次生成的复杂度。实验表明，相比一步到位地同时生成所有物体，此逐次生成策略在训练未见过的极端布局上具有更强的泛化性。例如，当物体数量远超训练集或位置分布很不寻常时，传统模型可能失败，而 IterInpaint 仍能逐个正确绘制。其原因为：一步生成的模型只能学习训练分布的统计，而逐步生成可以自适应每个物体的条件，不易受整体布局异常的干扰。当然，迭代方法也有代价，即生成速度变慢(需要多次扩散采样)且过程设计较复杂(如决定生成顺序、物体间遮挡处理等)。Cho 等人对生成顺序等因素进行了消融研究，发现合理的顺序和区域处理对结果影响很大。IterInpaint 反映出将扩散模型的全局生成任务分解为一系列局部子任务可以提高控制准确性，但需要权衡效率。

注意力引导方法是另一种无需改动模型就能控制布局的技术。其代表是英伟达提出的“Paint-With-Words”方法。该方法最早在 2022 年的 eDiff-I 模型[25]中被提出。思想是利用扩散模型中的跨注意力(cross-attention)机制：在文本到图像扩散模型中，文字提示经过 Transformer 编码生成 embedding，扩散 UNet 通过跨注意力将文本嵌入映射到图像特征上，从而将语义注入图像生成。如果我们希望某个单词对应的物体出现在图像特定位置，可以人为干预注意力图。具体操作是在扩散采样过程中，对于用户指定的“单词–图像位置”对应关系，提高该单词对相应图像区域像素的注意力权重。实现上，用户提供一张与布局类似的分割掩码：例如一个掩膜区域标注为“cat”，另一个区域标注为“tree”。在每次扩散迭代计算 cross-attention 时，检测对应于“cat”这个 token 的注意力矩阵，并将其在掩膜所指示的像素位置的权重调高(或在其它位置调低)，从而强制模型将“cat”相关的形状内容放入该区域。这种方法相当于引导模型注意去“在指定位置画出某物体”。由于不需要改模型，仅在推理时操作，它可以用于任何预训练文本扩散模型如 Stable Diffusion。实践证明，Paint-With-Words 能够纠正纯文本生成中物体错位或缺失的问题，使诸如“天空中一只鸟，旁边一棵高树”这类描述能够得到在正确位置包含这些元素的图像。这实际上是一种人为控制的布局引导，与我们本文讨论的自动布局到图像任务略有不同。但它启发了后续一些研究：例如最近有工作提出让扩散模型接受多段文本分别控制图像不同区域(称为多提示扩散)，以及多扩散 MultiDiffusion 方法通过在同一次采样中融合多个扩散条件输出来满足多个区域约束。总体而言，注意力操纵法的优点是无须训练、精度高，但需要用户输入更丰富的信息(例如区域掩膜)。在自动场景下，如果掩膜由算法预生成，也可用于提升布局控制，但预生成准确性会直接影响最终效果。

还有一种新颖的推理方法是由 Shirakawa 等人提出的 NoiseCollage(噪声拼接)方法[26]，将扩散模型

的噪声空间作为直接操纵对象，以实现布局控制。核心想法是：在扩散采样的初始阶段，为布局中的每个物体单独生成一片高斯噪声，然后根据物体的边界框将这些噪声片段裁剪并嵌入到全局噪声图的对应位置。之后，用这个“拼贴”而成的噪声图作为扩散采样的起点进行去噪。由于初始噪声已经在目标位置埋下了各物体的独立噪声模式，扩散过程自然会在这些区域生成相应物体，而不同区域的噪声互不影响，从而达到直接的布局控制。这种方法等价于将整体生成拆分成独立的局部生成再合并，但与迭代方法不同的是它在一个扩散过程内完成了这一合并，避免了多次连续采样。NoiseCollage 被称为“培训开销为零”，因为它完全利用了预训练扩散模型本身，无需训练额外网络即可使用。NoiseCollage 的提出印证了“噪声是良好的直接布局控制媒介”这一观点。它的优势在于不需要模型改动或训练，且能够避免一些注意力操纵方法可能引入的长程依赖干扰，相当于提供了一种简单直观的“用噪声表达布局”的范式。不过，该方法目前仍属新兴思路，其在高度复杂场景(如大量重叠或交互的物体)下的效果还有待深入研究。

除了上述工作外，Mixture of Diffusers [27]通过加权融合多个扩散路径的中间特征实现区域解耦控制；MultiDiffusion [28]提出时空联合的噪声调度策略，支持多提示条件在单一采样过程中的协同优化；RPG [29]采用分阶段递归生成，先构建场景拓扑骨架再逐步细化实例；DenseDiffusion [30]引入密集像素级条件反射机制，通过隐空间插值实现布局微调；Omost [25]则训练了专用大语言模型，根据提示词自动生成复杂的布局和详细的提示，最后在推理过程中动态调整注意力权重以满足复杂布局约束。综上所述，推理阶段的方法提供了灵活高效的替代手段，尤其在模型零样本泛化和用户交互方面具有优势。当我们无法为每种新布局都重新训练模型时，像迭代填充或噪声拼接这样的技术依然能够产出合理结果。它们的缺点主要在效率和使用便利性上：迭代方法比一次生成更慢，注意力操纵需要额外信息，噪声拼接需要一定先验假设。但随着计算加速和算法改进，这些方法完全可以与前述模型训练方法结合使用，为实际应用提供更强的控制能力。

4. 方法比较与分析

在方法优缺点对比上，我们可以归纳如下：专用扩散模型针对布局任务量身定制，因而在封闭集上的布局精度和图像质量最为卓越，其劣势是训练成本高且难以涵盖开放场景；基于预训练模型的方法充分利用大模型知识，取得了概念泛化和训练高效的优势，但需要权衡如何不损伤原模型性能，如 GLIGEN 采用冻结策略，LayoutDiffuse 引入残差适配。相较专用模型，它们对数据量依赖更低，但在极端布局下可能不如专用模型稳健，需要辅以推理技巧。推理控制方法灵活且训练成本为零，能按需应用于任意预训练扩散模型，特别适合用户交互式场景。然而，迭代方法在物体间关系复杂的情况下可能出现次序依赖问题(如先生成的物体可能影响后生成部分的背景连续性)，注意力引导则需要精细调节参数避免引入伪影，而噪声拼接目前主要支持矩形区域，对更复杂形状的物体布局支持有限。尽管如此，这些推理方法可以与模型改进方法互为补充：例如，将 GLIGEN 生成的初步图像再用 Paint-With-Words 细调局部，或在 LayoutDiffuse 基础上加入 NoiseCollage 的噪声初始化，以进一步强化布局满足程度。总结后的主流方法特性对比分析如表 1 所示。

Table 1. Comparative analysis of mainstream methods for layout-to-image generation
表 1. 主流方法特性对比分析

方法类型	核心优势	主要局限性
任务专用扩散模型	布局精度高，质量优	训练成本高，场景封闭
预训练适配方法	泛化性强，数据高效	依赖预训练知识迁移
零样本推理控制	零训练成本，高灵活性	计算效率低，需人工干预

5. 未来研究展望

与早期的生成对抗网络(GAN)方法相比，扩散模型在提升图像清晰度和逼真度的同时，也显著增强了对复杂场景中多物体布局的控制能力，其核心突破在于构建了多模态空间对齐的噪声预测范式。现有方法通过布局编码器、预训练适配器及推理引导策略，实现了从封闭场景到开放概念的可控生成，但仍然存在一些不足：目前大多数方法依赖于对空间逻辑的被动拟合，对动态交互、物理规律以及开放语义的建模尚未触及生成式人工智能的认知本质。最新研究表明，扩散模型正从“数据驱动生成”向“知识引导创造”演进。未来研究可聚焦于以下几个方向：1) 开放世界的因果生成架构：结合视觉语言大模型(VLMs)的开放语义理解与神经符号系统的逻辑推理能力，构建可解释的布局表征空间。借鉴 Lavin-DIT [31]的多模态联合编码框架，将边界框坐标扩展为包含物理属性(材质、运动轨迹)的语义张量，通过 IDM 可逆网络[32]实现视觉概念与空间参数的解耦控制。这不仅能支持“未见物体 + 未知位置”的零样本生成，还可通过 AF-LDM [33]的等变损失约束，确保生成结果符合刚体运动等物理规律。2) 动态场景的时空一致性建模：基于 Sora [34]模型的时空 patch 表征机制，将布局控制从静态平面延伸至四维时空(3D 空间 + 时序演化)。通过 MagicAnimate [35]的分段重叠帧预测技术，在扩散过程中同步优化物体运动轨迹与外观连续性，解决传统方法在动态场景中的闪烁伪影问题。可融合 DREAM-Talk [36]的表情驱动策略，构建基于注意力重定向的时序传播模块，实现跨帧布局的因果关联。3) 人机协同的闭环优化系统：设计“生成–评估–迭代”的交互范式，将 DesignGPT [37]的强化学习反馈机制与 VA-VAE 的潜在空间对齐技术结合。用户可通过自然语言指令实时调整布局参数(如“将树木向右移动并增加光影效果”)，系统利用 Attentive Eraser [38]的注意力抑制技术定位修改区域，结合 AF-LDM 的带限特征重构实现局部精细化编辑。这种交互范式将重构创作生产关系，使 AI 从执行工具进化为创意伙伴。

可以得出的结论是，扩散模型正在突破传统生成技术的工具属性，向具备空间认知能力的智能创造媒介转变。以 Lavin-DIT 展现的上下文学习能力为例，其已初步实现从简单的布局输入到复杂空间关系推理的跨越。展望未来，通过融合神经微分方程与符号逻辑系统，布局生成有望发展出类似人类的空间想象能力——不仅能够精确定位和放置物体，还能自主推理诸如“桌子应支撑花瓶”、“交通标志应立于路口”等隐性约束。这种从“坐标映射式工具生成”向“空间智能认知创造”范式的跃迁，或将重新定义 AIGC 的技术边界，使机器创造力迈向融入物理常识与价值判断的新阶段。

参考文献

[1] Cheng, J., Liang, X., Shi, X., *et al*. (2023) LayoutDiffuse: Adapting Foundational Diffusion Models for Layout-to-Image Generation. arXiv: 2302.08908. http://arxiv.org/abs/2302.08908

[2] Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y. and Li, X. (2023) Layoutdiffusion: Controllable Diffusion Model for Layout-To-Image Generation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 22490-22499. https://doi.org/10.1109/cvpr52729.2023.02154

[3] Cho, J., Li, L., Yang, Z., *et al*. (2024) Diagnostic Benchmark and Iterative Inpainting for Layout-Guided Image Generation. arXiv: 2304.06671. http://arxiv.org/abs/2304.06671

[4] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Networks. arXiv: 1406.2661. http://arxiv.org/abs/1406.2661

[5] Ashual, O. and Wolf, L. (2019) Specifying Object Attributes and Relations in Interactive Scene Generation. arXiv: 1909.05379. http://arxiv.org/abs/1909.05379

[6] Johnson, J., Gupta, A. and FEI-Fei, L. (2018) Image Generation from Scene Graphs. arXiv: 1804.01622. http://arxiv.org/abs/1804.01622

[7] Wang, B., Wu, T., Zhu, M., *et al*. (2022) Interactive Image Synthesis with Panoptic Layout Generation. arXiv: 2203.02104. http://arxiv.org/abs/2203.02104

[8] Sun, W. and Wu, T. (2021) Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis. arXiv: 2003.11571. http://arxiv.org/abs/2003.11571

[9] Arjovsky, M. and Bottou, L. (2017) Towards Principled Methods for Training Generative Adversarial Networks. arXiv: 1701.04862. http://arxiv.org/abs/1701.04862

[10] Radford, A., Metz, L. and Chintala, S. (2016) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv: 1511.06434. http://arxiv.org/abs/1511.06434

[11] Ho, J., Jain, A. and Abbeel, P. (2020) Denoising Diffusion Probabilistic Models. arXiv: 2006.11239. http://arxiv.org/abs/2006.11239

[12] Dhariwal, P. and Nichol, A. (2021) Diffusion Models Beat GANs on Image Synthesis. arXiv: 2105.05233. http://arxiv.org/abs/2105.05233

[13] Zhao, B., Meng, L., Yin, W., *et al*. (2019) Image Generation from Layout. arXiv: 1811.11389. http://arxiv.org/abs/1811.11389

[14] Kingma, D.P. and Welling, M. (2014) Auto-Encoding Variational Bayes. arXiv: 1312.6114. http://arxiv.org/abs/1312.6114

[15] Sun, W. and Wu, T. (2019) Image Synthesis from Reconfigurable Layout and Style. arXiv: 1908.07500. http://arxiv.org/abs/1908.07500

[16] Liang, J., Pei, W. and Lu, F. (2022) Layout-Bridging Text-to-Image Synthesis. arXiv: 2208.06162. http://arxiv.org/abs/2208.06162

[17] Perera, P., Nallapati, R. and Xiang, B. (2019) OCGAN: One-Class Novelty Detection Using Gans with Constrained Latent Representations. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 8576-8585. https://doi.org/10.1109/cvpr.2019.00301

[18] Liu, L., Ren, Y., Lin, Z., *et al*. (2022) Pseudo Numerical Methods for Diffusion Models on Manifolds. arXiv: 2202.09778. http://arxiv.org/abs/2202.09778

[19] Song, J., Meng, C. and Ermon, S. (2022) Denoising Diffusion Implicit Models. arXiv: 2010.02502. http://arxiv.org/abs/2010.02502

[20] Ho, J. and Salimans, T. (2022) Classifier-Free Diffusion Guidance. arXiv: 2207.12598. http://arxiv.org/abs/2207.12598

[21] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., *et al*. (2023) GLIGEN: Open-Set Grounded Text-To-Image Generation. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 22511-22521. https://doi.org/10.1109/cvpr52729.2023.02156

[22] Zhang, L., Rao, A. and Agrawala, M. (2023) Adding Conditional Control to Text-to-Image Diffusion Models. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3813-3824. https://doi.org/10.1109/iccv51070.2023.00355

[23] Wang, X., Darrell, T., Rambhatla, S.S., *et al*. (2024) InstanceDiffusion: Instance-Level Control for Image Generation. arXiv: 2402.03290. http://arxiv.org/abs/2402.03290

[24] Wang, X., Fu, S., Huang, Q., *et al*. (2025) MS-Diffusion: Multi-Subject Zero-Shot Image Personalization with Layout Guidance. arXiv: 2406.07209. http://arxiv.org/abs/2406.07209

[25] Balaji, Y., Nah, S., Huang, X., *et al*. (2023) eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv: 2211.01324. http://arxiv.org/abs/2211.01324

[26] Shirakawa, T. and Uchida, S. (2024) NoiseCollage: A Layout-Aware Text-to-Image Diffusion Model Based on Noise Cropping and Merging. arXiv: 2403.03485. http://arxiv.org/abs/2403.03485

[27] Jiménez, Á.B. (2023) Mixture of Diffusers for Scene Composition and High Resolution Image Generation. arXiv: 2302.02412. http://arxiv.org/abs/2302.02412

[28] Bar-Tal, O., Yariv, L., Lipman, Y., *et al*. (2023) MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. arXiv: 2302.08113. http://arxiv.org/abs/2302.08113

[29] Yang, L., Yu, Z., Meng, C., *et al*. (2024) Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. arXiv: 2401.11708. http://arxiv.org/abs/2401.11708

[30] Kim, Y., Lee, J., Kim, J., Ha, J. and Zhu, J. (2023) Dense Text-To-Image Generation with Attention Modulation. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 7667-7677. https://doi.org/10.1109/iccv51070.2023.00708

[31] Wang, Z., Xia, X., Chen, R., *et al*. (2025) LaVin-DiT: Large Vision Diffusion Transformer. arXiv: 2411.11505. http://arxiv.org/abs/2411.11505

[32] Chen, B., Zhang, Z., Li, W., *et al*. (2025) Invertible Diffusion Models for Compressed Sensing. arXiv: 2403.17006. http://arxiv.org/abs/2403.17006

[33] Zhou, Y., Xiao, Z., Yang, S., *et al*. (2025) Alias-Free Latent Diffusion Models: Improving Fractional Shift Equivariance of Diffusion Latent Space. arXiv: 2503.09419. http://arxiv.org/abs/2503.09419

[34] Liu, Y., Zhang, K., Li, Y., *et al*. (2024) Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv: 2402.17177. http://arxiv.org/abs/2402.17177

[35] Xu, Z., Zhang, J., Liew, J.H., *et al*. (2023) MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. arXiv: 2311.16498. http://arxiv.org/abs/2311.16498

[36] Zhang, C., Wang, C., Zhang, J., *et al*. (2023) DREAM-Talk: Diffusion-Based Realistic Emotional Audio-Driven Method for Single Image Talking Face Generation. arXiv: 2312.13578. http://arxiv.org/abs/2312.13578

[37] Ding, S., Chen, X., Fang, Y., *et al*. (2023) DesignGPT: Multi-Agent Collaboration in Design. arXiv: 2311.11591. http://arxiv.org/abs/2311.11591

[38] Sun, W., Cui, B., Dong, X.M., *et al*. (2025) Attentive Eraser: Unleashing Diffusion Model's Object Removal Potential via Self-Attention Redirection Guidance. arXiv: 2412.12974. http://arxiv.org/abs/2412.12974