引用格式：张新长,赵元,齐霁,等.基于AI大模型的文生图技术方法研究及应用[J].地球信息科学学报,2025,27(1):10-26. [Zhang X C, Zhao Y, Qi J, et al. Research and application of text-to-image technology based on AI foundation models[J]. Journal of Geo-information Science, 2025,27(1):10-26.] **DOI:**10.12082/dqxxkx.2025.240657; **CSTR:**32074.14.dqxxkx.2025.240657

基于AI大模型的文生图技术方法研究及应用

张新长[1,2,3],赵　元[4],齐　霁[2,3*],冯炜明[4]

1. 新疆大学地理与遥感科学学院,乌鲁木齐 830017; 2. 广州大学地理科学与遥感学院,广州 510006; 3. 广州大学黄埔研究院,广州 510000; 4. 广东省城乡规划建设智能服务工程技术研究中心,广州 511300

Research and Application of Text-to-Image Technology Based on AI Foundation Models

ZHANG Xinchang[1,2,3], ZHAO Yuan[4], QI Ji[2,3*], FENG Weiming[4]

1. The College of Geography and Remote Sensing Sciences, Xinjiang University, Urumqi 830017, China; 2. School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China; 3. Huangpu Research School of Guangzhou University, Guangzhou 510000, China; 4. Guangdong Urban and Rural Planning and Construction Intelligent Service Engineering Technology Research Center, Guangzhou 511300, China

Abstract: [Objectives] To systematically review recent advancements in text-to-image generation technology driven by large-scale AI models and explore its potential applications in urban and rural planning. **[Discussion]** This study provides a comprehensive review of the development of text-to-image generation technology from the perspectives of training datasets, model architectures, and evaluation methods, highlighting the key factors contributing to its success. While this technology has achieved remarkable progress in general computer science, its application in urban and rural planning remains constrained by several critical challenges. These include the lack of high-quality domain-specific data, limited controllability and reliability of generated content, and the absence of constraints informed by geoscience expertise. To address these challenges, this paper proposes several research strategies, including domain-specific data augmentation techniques, text-to-image generation models enhanced with spatial information through instruction-based extensions, and locally editable models guided by induced layouts. Furthermore, through multiple case studies, the paper demonstrates the value and potential of text-to-image generation technology in facilitating innovative practices in urban and rural planning and design. **[Prospect]** With continued technological advancements and interdisciplinary integration, text-to-image generation technology holds promise as a significant driver of innovation in urban and rural planning and design. It is expected to support more efficient and intelligent design practices, paving the way for groundbreaking applications in this field.

Key words: generative AI; AIGC; image generation; text-to-image; diffusion model; artificial intelligence;

收稿日期：2024-11-27；**修回日期：**2024-12-14.

基金项目：国家自然科学基金面上项目(42071441、42371406)。 [**Foundation items:** National Natural Science Foundation of China, No.42071441, No.42371406.]

作者简介：张新长(1957—),男,新疆乌鲁木齐人,博士,教授,国际欧亚科学院院士、俄罗斯工程院外籍院士,主要从事空间数据整合及自适应更新技术方法、数字城市(智慧城市)理论与方法、深度学习与自然资源要素分类和提取等方面的教学与研究工作。E-mail: zhangxc@gzhu.edu.cn

***通讯作者：**齐　霁(1995—),男,广东佛山人,博士,博士后,主要从事遥感影像智能理解等方面研究。E-mail: jameschi95@foxmail.com

foundation model

***Corresponding author:** QI Ji，E-mail: jameschi95@foxmail.com

摘要：【目的】系统梳理基于AI大模型的文生图技术的进展，并探讨该技术在城乡规划领域的应用。**【探讨】**本文首先分别从训练数据集、模型和评价方法3个视角出发，对文生图技术发展进行了全面、系统的回顾，以揭示其成功背后的推动性因素。尽管文生图技术在通用计算机领域取得显著进展，但城乡规划领域的实际应用中仍面临诸多关键挑战，包括缺乏高质量领域数据、生成内容的可控性和可靠性不足，以及缺乏地学先验知识约束等。针对这些问题，本文提出了相应的研究思路，包括：面向领域需求的文生图数据增强策略、基于指令拓展的空间信息增强文生图模型、以及基于诱导布局的局部编辑文生图模型。在此基础上，结合多个实际应用案例展示文生图技术在城乡规划设计领域的应用价值和潜力。**【展望】**文生图技术通过技术突破和多学科融合，有望成为城乡规划设计领域的重要创新动力，为高效、智能化的设计实践提供支持。

关键词：生成式AI；AIGC；图像生成；文生图；扩散模型；人工智能；大模型

1　引言

近年来，人工智能（Artificial Intelligence，AI）技术取得飞速发展，主要可分为判别式和生成式两大类[1]。其中，判别式AI（Discriminative AI）的核心目标是学习输入数据X与输出标签Y之间的直接映射关系，即条件概率分布$P(Y|X)$，从而在给定输入的条件下预测合适的输出结果。如图1（a）所示，判别式AI主要旨在解决分类、目标检测等判别型任务。生成式AI（Generative AI）不仅关注输入与输出之间的关系，还重在理解输入数据本身，即学习数据的联合概率分布$P(X, Y)$或输入数据自身分布$P(X)$，从而能生成新的数据样本。如图1（b）所示，生成式AI致力于解决文本生成、图像生成等生成型任务。通过对比两者在核心目标和建模方式上的根本差异可以发现，判别式AI由于仅关注“输入–输出”之间的映射关系，往往容易忽视对数据本身进行深层次理解。而生成式AI必须基于对数据分布及底层结构的深入理解，才能有效生成符合人类预期的结果。从技术角度看，生成式AI在解决高维度、开放性和复杂性问题上，展现出卓越潜力。

2023年Gartner在研究报告①中就将生成式AI定义为未来的战略技术。目前，生成式AI已为科学研究[2-4]、艺术与设计[5-6]、教育[7-9]、工业生产[10]、商业[11-12]等众多领域带来了深刻变革[13]。在生命科学研究领域，AlphaFold系列生成式AI模型使蛋白质结构预测变得前所未有地高效、廉价和普及。通过减少对传统耗时且高成本手工实验的依赖，AlphaFold显著加速了药物研发、疾病研究和生物制品设计进程[3,14]。AI的突破甚至引发了科学研究的范式转变[15]。在地学领域，中国科学院发布了全球首个多模态地理科学大模型“坤元”（Sigma Geography）②。与ChatGPT等通用领域的大模型不同，坤元大模型根据地学领域面临的问题和需求进行设计，具备丰富的地理专业知识，从而能能实现地理专业问题解答、专业文献智能分析、地理数据资源

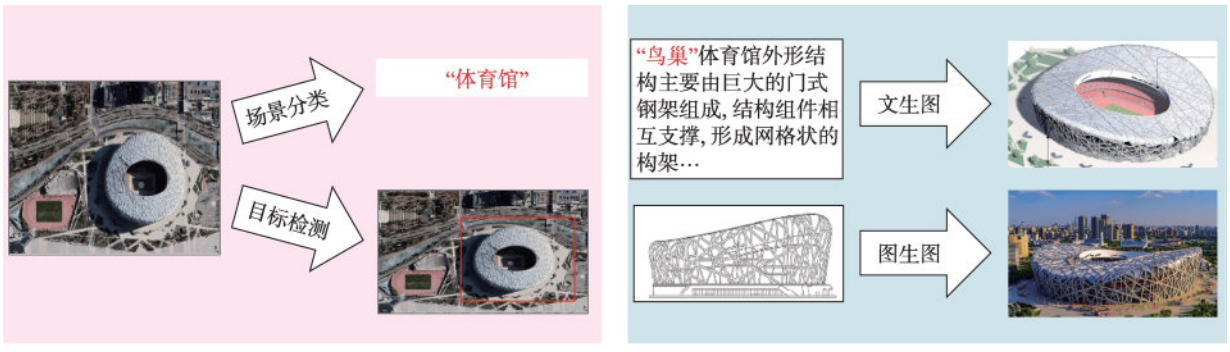


图1　判别式与生成式人工智能

Fig. 1　Discriminative vs. generative artificial intelligence

① https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises.

② https://igsnrr.cas.cn/news/picnews/202409/t20240920_7374237.html.

查询与分析、专题地图绘制等一系列功能。

生成式AI模型为各行业垂直领域带来重大机遇，但也面临诸多挑战。其中一个关键问题是如何将行业特定任务需求传达给模型，来激发其相关“记忆”[16-18]。对于已经从海量训练数据中获取了丰富知识的生成式AI大模型而言，恰当的“提示”是充分发挥其创作性潜能的关键，更是确保其生成内容真正满足用户预期的前提[19]。语言文本作为人类表达意图的基本媒介，凭借其丰富的语义和便捷性，是人与AI模型“沟通”的首选。因此，基于文本引导AI模型进行图像生成或编辑的“文生图”（Text-to-Image）技术，长期备受学术界和工业界的广泛关注[20-21]。这种将语言描述高效转换为图像或视频等数字化内容的技术，在许多行业中都展现出巨大的应用潜力，包括但不限于游戏开发、广告创意、城乡规划设计等专业领域。

在大数据、大模型的共同加持下，Midjourney、DALL-E[5,22]、Stable Diffusion[23-24]等代表性生成式AI在文生图技术上持续取得突破性进展。它们能以极高的效率生成高质量的图片甚至视频内容，显著降低艺术与设计的门槛，使许多非专业人士也能参与创作，甚至从中获利[25]。然而，这些文生图方法在解决特定行业的个性化需求时仍存在明显不足。这一现象在城乡规划设计领域中尤为突出。具体而言，地理数据的高复杂性和对空间结构的特定要求，使得现有模型很难直接应用。此外，文本提示的有效性也受到局限，往往无法充分表达城乡规划设计领域的复杂概念，从而影响了生成图像的质量和专业性。总体而言，如何将前沿的文生图技术转化为行业实际生产力，依然存在诸多严峻挑战。

在此背景下，本文首先对文生图数据进展进行了全面且系统的梳理，以揭示其背后的关键性推动因素；然后分析文生图技术在城乡规划设计行业应用落地所面临的问题和潜在解决思路；接着介绍了文生图技术在城乡规划设计领域的实际应用案例，包括老旧小区改造、工业区规划图生成、以及乡村局部改造规划设计等；最后对文生图技术未来发展进行展望。

2 文生图技术的发展历程

俗语“一图胜过千言万语”揭示了图像在信息表达浓缩性方面的独特优势。这说明，实现文本到图像的自动生成不仅是技术上的重要突破，更是迈向通用人工智能的关键一步[26-27]。以下从数据、模型和评价指标三个视角出发，对文生图技术进行回顾，以理解其背后的关键性推动因素。

2.1 文生图数据集

生成式AI的成功依赖于大规模、多样化的数据集。文生图（Text-to-Image）技术不仅追求图像生成质量，还要求生成结果与输入文本提示在相关性和语义上一致。因此，训练数据集不仅需具备规模和多样性，还要注重图像-文本描述的准确性与精细度。

首先，训练数据集的规模和多样性直接决定文生图模型的生成潜力。如果样本数量十分有限，或在内容、风格、主题上缺乏多样性，模型生成的图像会趋于单一，难以满足多样化创作需求。并且，样本均衡性也会影响模型实际表现。当某些类型的图像在数据集中占比过高时，模型倾向于生成该类型的图像，限制创造力。相反，主题、风格和场景均衡多样的数据集有助于模型学到更丰富的视觉概念，从而提升不同情境下的生成能力。

其次，文本描述的质量对文生图模型生成符合用户预期的图像至关重要，主要体现在2个方面（图2）。

（1）文本描述的详细程度通常影响模型对文本提示的理解与响应能力。高质量的文本描述不仅包括图像场景的整体语义或关键对象类别，还涵盖对象的空间分布、属性信息（如大小、形状、颜色等）以及对象与背景的关系信息。理论上，更详细的文本描述有助于模型具备更强的语义理解能力，从而响应更复杂的文本提示。然而，详细的文本描述意味着高昂的标注成本。为降低标注成本，研究人员已探索自动化或半自动化的图像-文本数据集构建方法。例如，OpenAI的Radford等[28]通过互联网爬虫构建了包含4亿图像-文本样本的数据集。基于类似方法，谷歌的Jia等人构建了包含18亿图像-文本样本的数据集ALIGN[29]，但并未开源。为便于更多研究人员训练文生图AI大模型，AI开源组织LAION与9所高校或机构合作，发布了58.5亿样本的LAION-5B数据集[29]。为减少错误，LAION-5B构建过程借助视觉-语言大模型CLIP[28]自动评估图文相似度，过滤不匹配的图文对。在地学领域，因数据来源的特殊性，如此大规模的图文数据集难以通过网络爬虫获取。对此，Zhang等[30]使用高性能

图2　文生图训练数据集质量的内涵

Fig. 2　Two dimensions of the quality of the Text-to-Image training dataset

的视觉–语言大模型BLIP-2[31]和CLIP，为遥感影像生成文本描述，构建了500万样本的RS5M数据集，以促进学领域的相关研究。

（2）文本描述与图像的关联粒度影响模型掌握文本–图像映射关系的能力。根据图文关联的精细程度，数据集可分为：场景级[28,32-33]、对象级[34-35]和像素级关联[36-38]三大类。场景级关联数据集提供文本标签或描述与图像整体的粗略对应关系。对象级关联数据集对关键对象进行了详细标注，提供更细粒度的文本-图像对应关系。像素级关联数据集对图像进行像素级标注，明确像素区域与文本之间的对应关系。随着关联粒度提升，模型能更容易理解视觉对象与文本提示的语义映射关系，进而实现精细的文本到图像转换。但如果训练数据中的文本描述不准确、不完整或与图像不匹配，可能导致模型学习错误的“文本–图像”对应关系，难以生成符合预期的图像。

随着高质量图像–文本数据集的涌现，文生图技术不断进步，逐步满足多样化应用需求。然而，与数十亿样本规模的通用图文数据集相比，地学领域图文数据集的规模和多样性明显不足。目前，最大的遥感影像–文本数据集RS5M仅为百万级别。同时，数据标注过程复杂且依赖专业知识，现有地学领域图文数据在文本描述的详细程度和图文关联粒度上仍有较大提升空间。因此，要推动文生图技术在特定领域的发展，亟需构建高质量、专业化的图像–文本数据集，以满足实际应用需求。

2.2　文生图模型

不同时期的代表性文生图模型如图3所示。其中，AlignDRAW作为文本生成图像领域的开创性研究，虽然奠定方法论基础，但生成的图像通常粗糙且细节欠佳。为提高图像质量及图文语义一致性，技术不断迭代，催生了一系列更先进的文生图模型。整个过程可大致可以分为以下4个关键阶段。

2.2.1　启航阶段：生成对抗网络的奠基作用

Goodfellow等[39]于2014年提出生成对抗网络（Generative Adversarial Networks, GANs），为图像

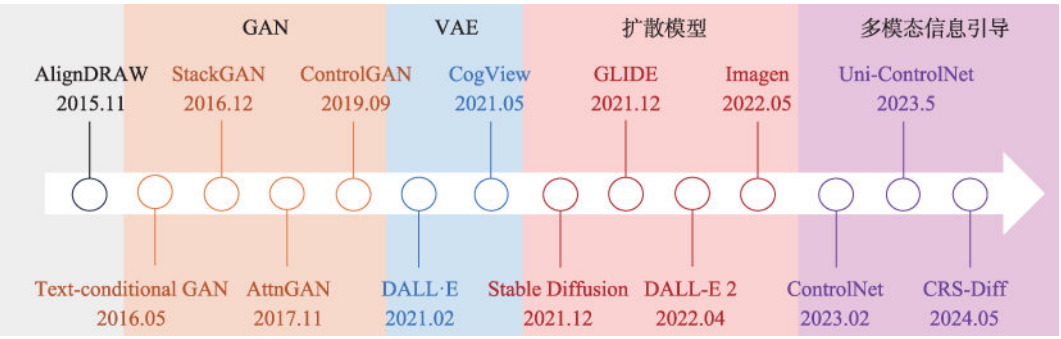


图3 不同时期的代表性文生图模型：基于GAN、VAE、扩散模型和多模态信息引导的方法

Fig.3 Representative text-to-image models over time: Including GAN-based, VAE-based, diffusion model-based and multi-modal information guided methods

生成研究奠定了良好基础。GAN由生成器和判别器组成，生成器通过生成逼真图像来欺骗判别器，判别器则区分真实图像和生成图像[40-41]。对抗训练方式帮助模型逐步学习真实图像分布，从而提升生成结果的真实性。然而，此类模型通常面临训练不稳定和模式坍塌（即生成大量重复且无意义的数据）等问题，难以生成高质量图像[42]。为此，许多学者在模型结构[43]、训练策略[44]等方面对GAN模型进行改进[20,45-46]。尽管此类研究取得一定进展，但仍未解决GAN训练中的两大深层原因：① 生成器与判别器能力失衡导致的梯度消失；② 传统GAN模型生成器损失函数隐含的基本假设“模型生成分布与真实图像分布部分重叠”，但实际训练中（尤其是初期）该假设难以优化，最终引发模式坍塌。针对这些问题，Gulrajani等[47]从GAN基本原理出发，提出用平滑且可导的Wasserstein距离代替Jensen-Shannon散度，从根本上改善了梯度消失和模式坍塌问题，最终提升了图像生成质量。

为实现可控图像生成，GAN的重要变体——条件GAN（Conditional GAN，CGAN）[48]通过在生成器或判别器中引入类别标签，指导从随机噪声到图像的生成，进而得到符合预期的特定结果。此后，Reed等[49]提出Text-conditional GAN，实现了文本提示引导的图像生成。然而，在实际应用中，基于GAN的文生图方法常常难以取得理想效果，主要有以下3个原因：① 缺乏语义解码与理解机制，仅能响应部分简单的类别标签，难以理解复杂文本描述；② 缺乏有效的图像和文本交互机制，导致图文对齐困难，生成结果难以控制；③ 生成器受随机噪声影响较大，即便加入条件信息，生成结果仍具有不可预测性。

总体而言，GAN及其变体CGAN为文生图技术启航奠定了基础[48,50]，但其网络结构和学习机制的不足导致早期图像生成质量和图文一致性难以提升。

2.2.2 发展阶段：变分自编码器的引入与发展

变分自编码器（Variational Auto-encoders，VAE）是典型的深度生成模型，其网络架构与传统自编码器一致，由编码器与解码器构成。但传统自编码器旨在学习确定性映射，将高维数据x编码为低维隐变量z，并对z解码重建原始数据。VAE则旨在学习隐变量z的概率分布，以提高生成数据的质量和多样性。VAE的核心工作流程为：① 编码器将输入数据x映射至潜在空间，输出z的概率分布参数（如均值μ和标准差σ）；② 通过重参数化技巧从概率分布中采样z，并将其输入解码器中以生成重构数据x'。训练VAE的目标函数包括两部分：重构损失$-\mathbb{E}_{q(z|x)}[\log p(x|z)]$和KL散度损失$KL(q(z|x)//p(z))$。重构误差保证生成数据与原始数据的相似性，KL散度使隐变量z接近预设先验分布（通常为标准高斯分布），以得到结构化的潜在空间，从而为可控生成奠定基础。

条件VAE（Conditional VAE，CVAE）[51-52]通过将文本标签信息作为额外条件信息来指导生成过程，最终实现基于文本引导的可控生成。与GAN相比，VAE在文生图任务中的关键优势在于：① 训练稳定性：重参数化技巧提升训练稳定性，减少模式坍塌风险；② 生成可控性：潜在空间具备良好结构性，可通过调整隐变量z生成多样化结果。然而，传统VAE和CVAE在实际应用中效果有限[53]。主要原因在于KL散度的局限性与计算复杂性，以及隐变量z后验分布假设过于理想化。因此，后续研究

从损失函数、后验分布假设和模型架构入手改进VAE。其中，向量量化变分自编码器（Vector quantised VAE，VQ-VAE）[54-55]，通过离散化的潜在空间向量建模方式，解决传统VAE在捕获高质量离散结构上的不足。这一设计显著提升图像生成质量与可控性潜力。基于VQ-VAE，OpenAI于2021年提出DALL-E[5]，实现文生图领域的革命性突破，其成功的关键在于：

（1）超大规模高质量数据：OpenAI构建约2.5亿对图文训练DALL-E，使其广泛学习各种语义对象和场景组合，提升生成能力。

（2）任务分解与分步训练：DALL-E将文生图任务拆解为3步：独立编解码、文本到图像映射和图文匹配优化，并逐步训练各个子模块以协同提升文生图效果。

（3）模型结构改进：使用Transformer建模文本–图像联合分布，提升建模效果、并行效率和拓展性。

（4）引入CLIP模型增强图文匹配度：DALL-E用预训练的CLIP模型对文本和生成图像进行匹配评分与重排序，增强生成结果与文本提示的匹配度。

得益于DALL-E在图像生成质量和图文语义一致性上的突破，文生图技术迅速引发广发关注并进入快速发展阶段[56]。

2.2.3 爆发阶段：扩散模型的兴起与技术突破

扩散模型（Diffusion Models）是一种新兴的生成模型，其原理从根本上区别于以往的生成方法。它通过将复杂的图像生成过程分解为一系列简单的“去噪”步骤，在生成图像的质量和多样性上实现突破。具体实现主要为2个过程：① 前向扩散，对原始图像逐步添加噪声，直至完全变成高斯噪声。② 反向生成，模型逐步移除噪声，恢复图像细节信息。充分训练后，模型可掌握数据特征分布，并从随机噪声生成高质量新图像。尽管扩散模型潜力巨大，但研究初期却面临诸多挑战：

（1）生成效率低：经典扩散模型需多次迭代生成图像，效率远低于GAN和VAE。

（2）最大似然偏差：扩散模型在训练过程中通常优化的是近似变分下界，而非直接最大化真实对数似然，导致生成性能损失。

（3）数据泛化能力不足：扩散模型高度依赖训练数据，难以生成超出训练集分布以外的高质量样本。

针对上述挑战，后续研究在DDPM[57]、SGM[58]和Score SDE[59]等经典扩散模型基础上，设计了更高效的采样机制[60-62]和更精确的似然和密度估计[63-64]，并适配了更广泛的数据类型。此后，扩散模型逐步克服了早期局限，在图像生成质量上超越GAN和VAE模型，为图像生成技术的爆发式增长奠定基础。

在文生图领域，2021年Dhariwal和Nichol[65]提出类别引导的扩散模型，可接收类别信息指导图像生成。随后，OpenAI提出GLIDE，可生成与文本提示高度匹配的逼真图像，并支持语言引导下的图像全局或局部编辑。2022年，OpenAI推出DALL-E 2[22]，以强大且可控的图像生成能力和数字化内容创作潜力，将基于扩散模型的文生图技术引入公众视野。自此，文生图技术及其应用迎来爆发增长。Imagen、stable diffusion、以及Midjourney等强大的文生图模型相继涌现，推动文生图技术在各领域的应用。

2.2.4 深化应用：多模态信息引导的精细化生成控制

提升图像生成的精细化控制能力以满足用户的个性化需求，对推动文生图技术在艺术、电商、建筑设计、城乡规划等行业应用至关重要，但也极具挑战。在此背景下，单一文本信息引导图像生成存在两大局限：

（1）文本的模糊和多义性：文本语言高度抽象，难以精准描述目标和场景细节，且词句的多种含义增加生成的不确定性。

（2）文本表达空间和结构信息的能力有限：建筑设计和城乡规划等任务涉及复杂空间关系和结构信息，难以通过文本直接描述。

针对上述问题，近期研究[66-67]旨在建立多模态信息协同引导的生成控制机制，支持文本描述、边缘轮廓图、语义分割标签、深度估计图、姿态估计图等不同模态的提示，实现精细化控制。目前，多模态引导的文生图方法已展现出更广阔的应用前景。

2.2.5 文生图模型小结

总结而言，文生图技术的迅速发展离不开生成式AI模型的不断进步。生成对抗网络（GANs）为图像智能生成奠定了基础。变分自编码器（VAE）及其变体模型拓展了基于文本的生成能力。OpenAI推出的DALL-E和CLIP模型深化了文本与图像的语义关联，使生成图像能更精确地匹配文本提示。扩散模型的提出大幅提升了图像质量与多样性，推动文生图技术的爆发式增长。近期研究利用多模态信息引导图像生成，进一步提升文生图的可控性，推动文生图技术在多个领域的应用。

2.3 文生图技术评价方法

图像质量和文本–图像对齐度是评价文生图技术的两大维度。前者关注生成图像的真实性和视觉效果，后者衡量图像与文本描述的匹配度。

在图像质量评价中，常用指标之一是Fréchet Inception Distance（FID）[68]。FID计算生成图像分布与真实图像分布在高维特征空间中的Fréchet距离（又称Wasserstein-2距离），以评估图像质量。FID值越小，表明生成图像的分布越接近真实图像分布，图像质量越高。该指标简单可靠，具有明确的物理意义，因此广泛应用于图像生成任务。另一常用指标Inception Score（IS）[42]在视觉质量基础上，进一步考虑了生成结果的多样性。IS通过计算条件类别分布与边缘类别分布之间的KL散度，评估生成结果的多样性。IS分数越高，表明生成图像不仅视觉质量高，还有良好的类别多样性。

衡量图像与文本匹配度的代表性方法是利用预训练的视觉–语言模型CLIP计算图像–文本之间的余弦相似度（即CLIP Score）[69]。CLIP Score的值越高，表明生成图像与文本的匹配度越高。此外，R-Precision[70]通过计算生成图像与文本描述的特征余弦相似度，衡量生成图像与对应文本描述之间的语义一致性。R-Precision值越高，表示生成图像与真实文本描述之间的匹配程度越高。

除上述指标外，许多研究提出了专门的文生图评估基准[71-73]。例如，Multi-Task Benchmark[73]设置32种任务来全面地评估文生图技术能力，并将任务分为3个难度级别。这些评价方法为文生图模型性能分析提供了量化依据，并能客观反映当前模型的不足，以指导后续改进。随着文生图技术的发展，未来可能会涌现更多针对特定场景的评价指标，以推动模型在多样性、真实性和图文匹配度等方面的不断提升。

3 文生图技术方法在城乡规划设计领域的研究与应用

3.1 传统城乡规划设计的需求与难点分析

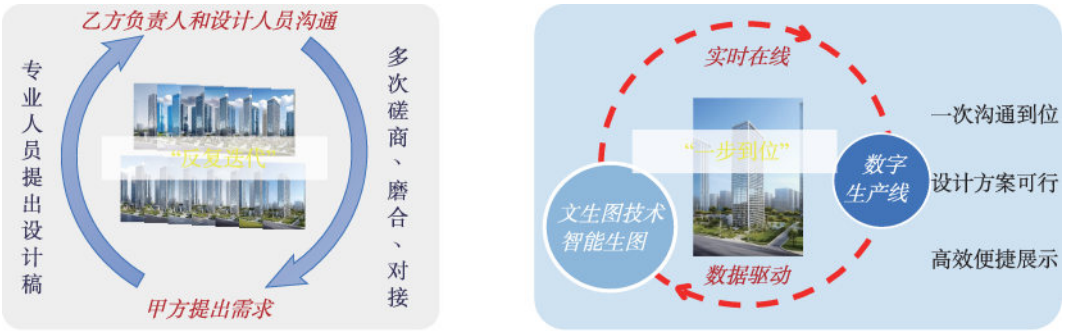
规划设计的核心任务是在结合现实条件的基础上，将甲方合理需求准确转化为专业设计方案。然而，在传统以人工为主的设计模式中，需求和设计双方往往需要耗费大量时间与精力，通过多次磋商和反复对接来明晰需求的内涵（图4(a)）。不仅显著增加了项目实施的时间和人力成本，还可能降低规划设计的质量。具体而言，传统设计模式下，双方沟通中面临以下难题：

（1）需求不明确：在项目初期，甲方可能尚未明确自身的部分隐性需求，甚至未能完全意识到这些需求。这需要设计团队通过深入调研或多轮沟通进行挖掘和确认。同时，甲方在初期的需求表达可能过于模糊或不完整，需要通过多次反馈与沟通逐步优化和完善。

（2）信息传递偏差：在沟通过程中，甲乙双方可能因表达方式或理解上的差异，导致信息传递失真。这种偏差需要通过持续的沟通和调整来纠正，方能确保设计方案准确体现甲方意图。

（3）重复性劳动：每轮沟通后，乙方需依据调整后的需求重新设计或修改方案，这往往涉及大量低效且耗时的重复性工作，降低了整体设计效率。

针对上述难题，亟需引入新的技术手段来优化沟通流程，提高项目推进的效率与实施质量。



(a) 传统模式下建筑规划设计　(b) 文生图模式下的建筑规划设计

图4 传统模式下的城乡规划设计与文生图模式下的城乡规划设计

Fig. 4 Urban and rural planning and design in the traditional paradigm vs. the text-to-image model

3.2 文生图技术为规划设计带来的机遇与挑战

文生图技术为城乡规划设计流程变革带来重要机遇（图4(b)）。文生图技术通过智能生成和数据驱动的数字化流程，能将甲方合理的文本描述高效转化为直观的规划设计方案，从而解决传统设计模式中繁琐的多轮沟通与反复修改问题。这种全新的设计范式有望显著提升设计效率、降低沟通成本。

从城乡规划设计流程（图5）来看，文生图技术在各个阶段的优化潜力包括：

（1）提升前期沟通效率：项目初期，甲乙双方通常需要反复沟通和磨合以确认需求。AI技术能对甲方简短模糊的需求描述进行详细解读。AI文本生成可显著缩短需求确认时间，降低沟通障碍。

（2）加快概念设计迭代效率：由于文本描述的抽象性，乙方在项目前期需根据甲方需求绘制概念设计图。然而，该过程通常需要反复修改，耗时耗力。对此，文生图技术能将文本需求快速转化为直观设计方案，供甲方反馈。这种即时生成技术显著减少了传统设计流程中的重复性劳动，加速设计方案的优化与更新。

（3）智能完善深化设计：在深化设计阶段，文生图技术可利用领域知识精细调整空间布局。例如，自动优化区域内的道路分布，提升空间配置的科学性。结合专业约束条件，文生图技术能生成符合行业规范的详细规划草案，提高设计方案的可行性。

（4）加速效果渲染：为将设计师创意完整且直观地呈现给客户，还需将精简线稿或三维模型渲染成实景效果图像。传统渲染方法往往需要耗费数小时，且依赖专业人员精心调校。而文生图技术能在几十秒甚至更短的时间内得到高质量的渲染结果，大幅提升效率。

文生图技术可为优化城乡规划设计流程提供新思路和技术途径，有望大幅提升设计效率与质量。但现有文生图技术在生成复杂设计方案时仍存在显著局限性。要将这一前沿技术转化为城乡规划设计领域的实际生产力，仍需克服以下关键技术与挑战。

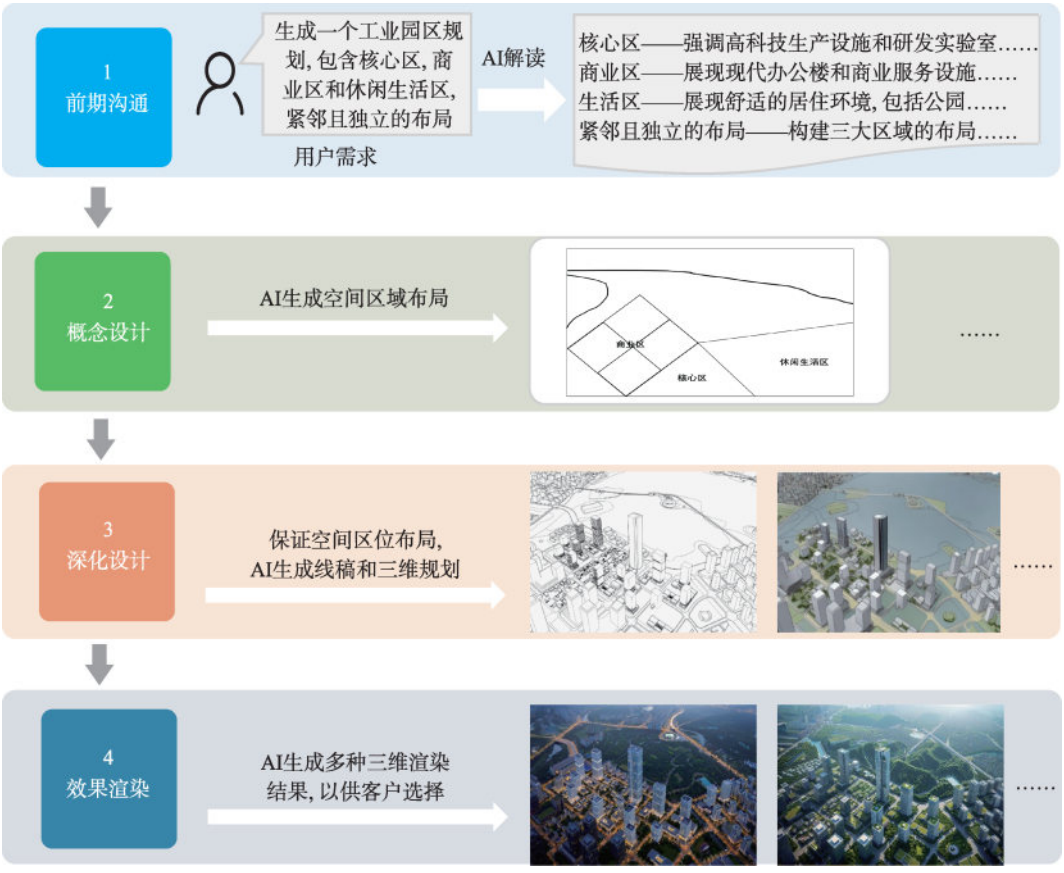


图5　文生图辅助下的城乡规划设计基本流程

Fig. 5　Basic process of urban and rural planning and design with the assistance of text-to-image methods

(1)缺乏高质量领域数据集

文生图模型的训练和优化高度依赖于大规模、高质量的专业数据集。然而,当前城乡规划设计领域的数据集普遍面临以下问题:① 样本匮乏:高分辨率图像、复杂地理场景的数据获取难度较高,且受隐私和版权限制;② 数据质量不足:公开数据集中标注不完整或不准确,且样本多样性差。这些问题制约了文生图技术在城乡规划设计中的实际应用。

(2)文本提示缺乏领域专业知识支撑

城乡规划设计领域的文生图任务通常依赖文本提示来描述场景目标。然而,目前的文本提示表达在以下方面存在不足:① 语义不明确:自然语言描述难以精确表达复杂的城乡规划设计概念,如“宜居性”、“空间效率”等;② 缺乏领域术语支持:当前文生图模型未能有效利用城乡规划设计中的专业术语和知识体系,导致生成结果与实际需求存在偏差;③ 文本-图像映射能力不足:模型对复杂场景中空间信息的映射能力有限,难以充分展现城乡规划设计领域的特定需求。因此,如何将城乡规划设计专业知识融入文本提示中,构建更专业化的城乡规划设计语言描述,是提升文生图模型生成质量的关键。

(3)生成过程缺乏领域先验知识约束

城乡规划设计领域的生成任务不仅追求视觉效果,还必须符合基本空间规律和设计规范。然而,现有文生图技术在生成过程中存在以下局限:① 缺乏先验知识引导:生成过程未能充分考虑功能区分布规律和城乡规划设计规范,导致生成无效方案;② 精准空间布局困难:现有技术难以处理复杂的空间布局,可能导致生成结果与预期不符。

综上,文生图技术在城乡规划设计中展现了极大的应用潜力,从高效沟通到智能生成,从快速迭代到优化布局,为设计流程带来了颠覆性改进。然而,特别是在城乡规划设计领域,文生图技术的实际应用仍面临着数据、表达与生成约束方面的挑战。未来的研究需进一步结合更丰富的城乡规划设计先验知识,以推动文生图技术在该领域的发展与应用落地。

3.3 面向城乡规划设计的文生图技术思路

本研究针对文生图技术在城乡规划设计中的挑战,提出以下技术思路:首先,针对数据不足的问题,设计了面向城乡规划领域的文生图数据增强策略,以提升模型对多样场景的适配能力。其次,为强化生成模型对空间信息的理解和控制,基于指令拓展方法开发了空间信息增强的文生图大模型。最后,提出一种基于诱导布局的局部编辑文生图大模型,旨在满足局部编辑与精准布局需求,能在复杂场景中实现精细化控制和动态调整。

3.3.1 面向城乡规划设计的文生图数据增强策略

文生图模型的高效生成能力依赖于高质量、多样化的训练数据。然而,当前城乡规划设计领域的规划设计仍普遍面临数据匮乏、样本分布单一以及语义信息不完整等问题。为解决这一瓶颈,通过数据增强策略扩充数据规模、提升数据质量,为模型提供更丰富的训练资源。文生图数据增强策略包括以下3个步骤:

(1)结合行业先验知识进行数据集优化

针对城乡规划设计相关文生图数据集不足的问题,充分结合行业经验与专业知识来提升粗糙的文生图数据的质量:① 粗糙文生图数据集建立:将城乡规划设计图输入视觉-语言大模型生成简单的低质量文本描述,并初步形成大规模的图像-文本对齐数据;② 行业语料知识库构建:广泛收集行业内书籍、项目材料、研究报告等相关文档,并利用文本识别技术建立行业语料知识库;③ 基于行业先验知识的文本描述增强:利用检索增强生成(Retrieval-Augmented Generation,RAG)技术来动态调取行业语料知识库的相关内容,来细化并完善步骤①的粗糙文本描述;④ 图文匹配度优化:对完善后图文数据进行质量评估与筛选,通过语义一致性评分(如CLIP评分)保留高质量图像-文本样本,进一步提升数据质量。

(2)文字引导的图像生成扩展

文字引导生成高质量图像数据是一种重要的数据增强策略:① 图像输入与语义生成:输入城乡规划设计图,利用大模型生成丰富的语义描述,形成图像到文本的语义关联关系;② 指令编辑与多样化生成:进一步基于语义描述,通过引导指令调整图像特性(如颜色、光影条件等),生成多样化的场景样本。

(3)基于文本结构的指令拓展

将城乡规划设计中的领域知识融入数据增强策略,可进一步提升数据的专业性和适用性:① 引入城乡规划设计先验知识:结合区域气候、地貌等信息生成更符合实际应用场景的数据样本;② 多层次奖励机制:在生成的图像中通过奖励机制筛选与文本描述高度相关的图像,确保数据符合城乡规

划的领域需求。

通过上述步骤，有望改善数据质量、规模和专业化水平，为推动文生图在城乡规划设计领域发展奠定基础。

3.3.2 基于指令拓展的空间信息增强文生图大模型

在城乡规划设计中，如何准确理解空间布局和表达复杂场景的逻辑关系是文生图技术面临的重要挑战。为此，探索了基于指令拓展的空间信息增强文生图大模型，即通过整合文本指令与空间信息，以提升模型在复杂布局场景中的生成精度与一致性。该模型的核心功能与工作流程主要为：

（1）空间信息的提取与编码：模型首先通过多模态大模型提取输入图像或草图中的关键空间信息，包括边缘特征、对象关系和轮廓布局。该阶段提取的低级视觉特征将被转换为高层次的空间语义描述，生成可用于指令拓展的结构化空间信息。

（2）文本细化与语义补充：在生成的初步空间语义基础上，模型结合文本指令，利用T5等语言生成模型对描述进行属性细化和补充。例如，针对“房屋旁有绿地”的描述，进一步生成“绿地包含灌木丛和乔木，步道铺设鹅卵石”等具体细节。

（3）空间语义与文本指令的深度融合：模型将细化后的文本与空间信息相结合，通过对空间布局逻辑的分析与优化生成最终的图像。指令与布局的深度融合有助于提高生成结果与文本提示的匹配度和空间布局合理性。

（4）图像生成与输出优化：最后，模型利用条件生成网络（如扩散模型）根据融合后的语义与空间信息生成高质量图像，实现输出图像在语义完整性与空间表达上具备更高一致性和可用性。

通过深度融合空间信息与文本指令，上述大模型在城乡规划设计领域具备以下潜在优势：① 生成布局更合理的复杂城乡场景，改善传统文生图技术在空间表达上的不足；② 通过语言模型的细化能力，提高生成图像的细节丰富性和语义一致性。③ 支持多输入形式（文本、草图等）的特性，使其在不同设计需求下具备更强的适应性。然而，该技术也存在一定局限。首先，其生成效果高度依赖于训练数据的多样性和质量，若数据不足或存在偏差，则可能导致生成图像空间信息不准确。同时，多模态模型的高维处理机制增加了计算成本，对硬件设备提出更高要求。

3.3.3 基于诱导布局的局部编辑文生图大模型

城乡规划设计中的场景布局通常包含多个功能区，这些区域在空间结构和视觉表现上需要高度协调。然而，传统文生图技术在处理局部区域编辑任务时，难以有效兼顾局部细节质量和整体布局协调性。为此，基于诱导布局的局部编辑文生图大模型通过分割布局、特征提取、多模态解读和精确控制等技术，实现对局部区域的高效编辑，同时保证整体场景的逻辑性和美观性。模型的核心功能与工作流程如下。

（1）精准区域分割与布局提取：模型首先通过分割模块对输入的场景图像进行功能区划分，并提取各区域的结构化特征。例如，在规划图中分离出住宅区、商业区、绿化区等功能区域，生成特定区域的空间特征层，为局部编辑奠定基础。

（2）多模态语义理解与布局优化：利用多模态模型对待编辑区域进行解读，从图像中提取视觉特征，并生成相应文本描述。这些描述可以进一步优化为高语义的空间布局指令，如“绿化区包含乔木、灌木与步道”。生成的布局指令为局部编辑提供了强有力的引导。

（3）智能局部编辑与生成优化：将区域特征与用户提供的文本指令相结合，模型通过扩展模型和控制模块（如ControlNet）生成局部的精细化结果。例如，用户输入“增加绿化覆盖率，将绿地面积扩大50%”，模型可在保持整体场景逻辑一致的基础上，生成绿化区域扩大的新规划图。通过条件约束与语义匹配，模型还可优化编辑后区域的细节呈现，如道路纹理、城乡材质等。

该模型可具备以下优势：① 精准区域编辑：可对特定功能区进行高精度局部编辑，满足用户的个性化设计需求。② 整体一致性保障：在局部编辑的同时，通过诱导布局确保整体场景逻辑的连贯性。

3.3.4 基于AI大模型的文生图技术框架

基于3.3.1—3.3.3节所述的文生图技术思路，我们结合AI大模型初步构建了面向城乡规划设计的文生图技术框架，如图6所示。该框架分为多个层级，自下而向上依次为：数据处理层、计算引擎层、模型推理与增强层、应用模型与策略、前端交互渲染和应用场景复杂以及前端呈现层。具体各层内涵如下。

（1）数据处理与存储层

该层包括原始数据的采集与标准化处理，涉及

图6 面向城乡规划设计的文生图技术框架

Fig.6 Framework of text-to-image technical for urban and rural planning and design

线稿图、建筑渲染图、规划设计图及专题图等多样化数据源。结合MySQL等关系数据库与Elasticsearch等分布式搜索工具，确保数据的高效存储与快速检索。同时引入Neo4j进行知识图谱构建，为语义关联提供支持。

(2)计算引擎与服务部署层

通过基于Linux-Ubuntu操作系统的计算引擎和服务部署环境，集成xInference与vLLM等高效推理框架，结合Redis实现任务缓存与调度优化。该层为后续复杂计算与模型推理提供了稳定的底层支持。

(3)模型推理与增强层

该层是框架的核心，基于扩散模型与大规模生成模型的结合，支持空间信息增强与文本指导图生成两种主要能力。利用强化学习优化生成策略，并通过智能规划算法与知识图谱嵌入实现模型输出的精准化与情境适配。

(4)应用策略层

针对城乡规划设计需求，设计了多种应用策略，包括数据增强生成策略、基于指令拓展的空间信息补全以及基于诱导布局的局部编辑生成。这些策略有效提升了生成结果的实用性和适应性。

(5)前端交互与渲染层

在用户端，基于React与TypeScript技术栈，结合Ant Design组件库，提供了高效的交互界面。通过Axios实现后端数据与前端的无缝通信，同时借助Recharts进行动态数据可视化，全面支持用户的设计与决策需求。

通过层次化的设计，该技术框架可有效整合AI大模型的生成能力、知识图谱的关联能力以及人机交互的友好性，有望提升城乡规划设计效率，提升规划设计的智能化、协同化与精准化。

4 城乡规划设计领域文生图技术应用案例及前景分析

基于3.3节所述的技术思路和AI大模型文生图框架，本节介绍了该技术在城乡规划设计领域的应用案例，并展望了其广泛的应用前景。

4.1 城乡规划设计领域文生图技术的典型应用案例

4.1.1 老旧小区改造规划设计

随着城市的快速发展，老旧小区改造成为迫切需求。文生图技术能高效生成多元化的设计方案，避免传统改造设计中常见的急功近利与大拆大建问题。该技术具备即时交互能力，可根据居民与改造团队的诉求，实时调整方案内容。此外，文生图技术在局部细节优化方面也展现出高度的灵活性与精准性。在图7所示的小区改造项目中，文生图技术能迅速生成多种风格的改造效果图，帮助居民直观对比和选择最契合实际需求的设计方案，实现技术与人文关怀的有效平衡。

图7　文生图技术服务于老旧小区改造规划设计

Fig. 7　Text-to-image technology for planning and designing the renovation of old neighborhoods

4.1.2 工业区规划设计

工业区规划涉及复杂的空间布局与功能区划分，传统设计方法往往因设计周期长、沟通成本高而效率低下。文生图技术通过将文本引导生成和指令渲染相结合，提供一种快速高效的解决方案。如图8所示，用户仅需输入文本描述，如“生成一个多功能工业园区建成效果示意图”，模型即可结合地形起伏、功能分区和视觉审美等条

图8　文生图技术支撑下的工业区建成效果示意图高效生成

Fig. 8　Efficient generation of schematic diagrams of industrial area supported by text-to-image technology

件，快速生成相应图像。这种方式不仅能提高设计效率，还便于用户直观感受和高效决策。

4.1.3 乡村局部更新改造规划设计

局部更新改造是推动乡村振兴的主要工作思路之一，但如何避免造破坏乡村整体风貌往往具有挑战性。对此，可在合适文本提示的引导下，利用文生图技术来对乡村全景图像进行局部编辑，精确设计村庄特定区域。用户可标记图像中需重新设计的区域，并给出文本要求（如“保持整体布局不变，对局部建筑区进行改造”），文生图模型即可生成改造效果图。文生图技术的这种局部设计能力在乡村社区改造、旅游村落规划等场景中具有广泛应用潜力。

4.2 城乡规划设计领域文生图技术前景展望

随着生成式人工智能技术的快速发展，文生图模型在城乡规划设计领域展现出极大的潜力，因具有“交互性、实时性、专业性”特点为城乡规划设计研究的创新提供了全新视角和工具支持。基于文生图技术的特点及其发展趋势（图9），可从以下4个方面展望其在城乡规划设计领域的应用前景。

顾问型的服务模式：文生图技术通过将自然语言转化为复杂的城乡规划设计图表，实现面向问题解决的智能化服务模式，为不同情景提供针对性的辅助决策支持。

贴心的沟通交流：文生图模型为规划设计过程提供高效的沟通途径。借助生成的高质量内容，设计者能够更加清晰地传递复杂信息，减少沟通障碍。

专家级的数字助手：文生图技术凭借强大的数据处理与生成能力，为城乡规划设计提供专家级辅助支持。设计师通过简短文字描述即可生成复杂的规划图纸、建筑草图或场地布局，显著提高设计效率与准确性。同时，结合大数据分析，文生图技术能自动进行场地分析、交通流量预测和环境影响评估，为设计方案提供科学依据。随着技术优化，文生图模型将通过深度学习提升专业性与个性化，提供定制化设计方案，成为规划师的重要助手。

放飞想象的翅膀：文生图技术不仅能够将现实需求转化为具体设计，还能突破传统设计思维的局限，激发出无限的创意空间。设计师可以通过与文生图模型的互动，快速实现各种构思的可视化，探索新的空间布局、功能组合或建筑形态，从而推动城乡规划设计的创新与突破，打破常规，开创更具未来感的设计方案。

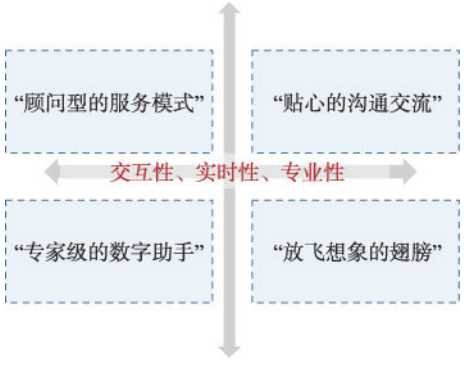


图9 文生图技术的未来愿景

Fig. 9 Future vision of text-to-image technology

5 结语

随着生成式人工智能技术的飞速发展，文生图技术为城乡规划设计带来了前所未有的机遇。它在设计流程优化、图像生成效率提升以及复杂场景建模等方面展现出显著优势。但目前相关研究仍处于探索阶段，且面临数据匮乏、未能充分利用领域先验知识、生成内容可控性不足等诸多挑战。需重点关注的问题有：

（1）专业化数据集构建：系统化收集与标注涵盖不同城乡风格、空间布局和功能区域的多样化数据，为模型训练提供坚实基础。

（2）知识与模型的深度融合：探索将城乡规划设计先验知识与生成模型深度结合，提升方案设计的科学合理性。

（3）可解释性与可靠性：加强文生图模型的可解释性研究，使生成结果的可信度，保障应用过程安全可靠。

（4）高效的多模态交互：探索多模态引导机制，使文生图模型能够灵活响应文本、图像、草图等多种提示信息，并生成更符合预期的设计方案。

文生图技术的快速发展正从根本上重塑城乡规划的设计范式。通过不断突破现有技术瓶颈，文生图技术将在更多复杂场景下展现出更强的创造

力和实用性，推动城乡设计与城乡规划设计应用迈向更高的智能化水平。

利益冲突：Conflicts of Interest

所有作者声明不存在利益冲突。

All authors disclose no relevant conflicts of interest.

作者贡献：Author Contributions

本研究由张新长构思；赵元和冯炜明完成实验设计和操作；张新长、齐霁、赵元和冯炜明完成论文的写作和修改。所有作者均阅读并同意最终稿件的提交。

The study was designed by ZHANG Xinchang. The experimental operation was completed by ZHAO Yuan, and FENG Weiming. The manuscript was drafted and revised by ZHANG Xinchang, QI Ji, ZHAO Yuan, and FENG Weiming. All the authors have read the last version of paper and consented for submission.

参考文献(References)：

[1] Tao C, Qi J, Guo M N, et al. Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5610426. DOI: 10.1109/TGRS.2023.3276853

[2] Wang H C, Fu T F, Du Y Q, et al. Scientific discovery in the age of artificial intelligence[J]. Nature, 2023, 620(7972):47-60. DOI:10.1038/s41586-023-06221-2

[3] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 [J]. Nature, 2024, 630(8016): 493-500. DOI: 10.1038/s41586-024-07487-w

[4] 张岸,朱俊锴.新一代人工智能驱动下地图学研究的机遇与挑战[J].地球信息科学学报,2024,26(1):35-45. [Zhang A, Zhu J K. Opportunities and challenges of cartography research driven by new generation artificial intelligence[J]. Journal of Geo-Information Science, 2024, 26(1):35-45.] DOI:10.12082/dqxxkx.2024.240128

[5] Ramesh A, Pavlov M, Goh G, et al. Zero-Shot text-to-image generation[C]//Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2021: 8821-8831.

[6] Epstein Z, Hertzmann A, the Investigators of Human Creativity, et al. Art and the science of generative AI[J]. Science, 2023, 380(6650): 1110-1111. DOI: 10.1126/science.adh4451

[7] Cooper G. Examining science education in ChatGPT: An exploratory study of generative artificial intelligence[J]. Journal of Science Education and Technology, 2023, 32(3):444-452. DOI:10.1007/s10956-023-10039-y

[8] Chiu T K F. The impact of Generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and midjourney[J]. Interactive Learning Environments, 2023: 1-17. DOI: 10.1080/10494820.2023.2253861

[9] Dan Y H, Lei Z K, Gu Y Y, et al. EduChat: A large-scale language model-based chatbot system for intelligent education[EB/OL]. 2023: 2308.02773. https://arxiv. org/abs/2308.02773v1.

[10] Rane N. ChatGPT and similar generative Artificial Intelligence (AI) for smart industry: Role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0 [J]. SSRN Electronic Journal, 2023. DOI: 10.2139/ssrn.4603234.

[11] Chen B Y, Wu Z X, Zhao R R. From fiction to fact: The growing role of generative AI in business and finance[J]. Journal of Chinese Economic and Business Studies, 2023, 21(4):471-496. DOI:10.1080/14765284.2023.2245279

[12] Kanbach D K, Heiduk L, Blueher G, et al. The GenAI is out of the bottle: Generative artificial intelligence from a business model innovation perspective[J]. Review of Managerial Science, 2024, 18(4): 1189-1220. DOI:10.1007/s11846-023-00696-z.

[13] 朱禹,叶继元.人工智能生成内容(AIGC)研究综述:国际进展与热点议题[J].信息与管理研究,2024,9(4):13-27. [Zhu Y, Ye J Y. A review of artificial intelligence generated content(AIGC): International progress and research agenda[J]. Journal of Information and Management, 2024, 9(4):13-27.]

[14] Callaway E. 'The entire protein universe': AI predicts shape of nearly every known protein[J]. Nature, 2022, 608(7921):15-16. DOI:10.1038/d41586-022-02083-2

[15] Li X, Guo Y L. Paradigm shifts from data-intensive science to robot scientists[J]. Science Bulletin, 2024 DOI: 10.1016/j.scib.2024.09.029

[16] Jia M L, Tang L M, Chen B C, et al. Visual prompt tuning [M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 709-727. DOI: 10.1007/978-3-031-19827-4_41

[17] Liu P F, Yuan W Z, Fu J L, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023,55(9):1-35. DOI:10.1145/3560815

[18] 刘安安,苏育挺,王岚君,等. AIGC视觉内容生成与溯源研究进展[J]. 中国图象图形学报,2024,29(6):1535-1554. [Liu A, Su Y T, Wang L J, et al. Review on the progress of the AIGC visual content generation and traceability[J]. Journal of Image and Graphics, 2024,29(6):1535-1554.] DOI:10.11834/jig.240003

[19] 王常圣. 面向大模型艺术图像生成的提示词工程研究[J]. 图学学报, 2024: 1-14. [Wang C S. Research on prompt engineering for large model art image generation [J]. Journal of Graphics, 2024:1-14.]

[20] 赖丽娜,米瑜,周龙龙,等. 生成对抗网络与文本图像生成方法综述[J]. 计算机工程与应用,2023,59(19):21-39. [Lai L N, Mi Y, Zhou L L, et al. Survey about generative adversarial network and text-to-image synthesis[J]. Computer Engineering and Applications, 2023,59(19):21-39.] DOI:10.3778/j.issn.1002-8331.2211-0392

[21] Zhang C S, Zhang C N, Zhang M C, et al. Text-to-image diffusion models in generative AI: A survey[EB/OL]. 2023:2303.07909.https://arxiv.org/abs/2303.07909v3

[22] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. 2022: 2204.06125.http://arxiv.org/abs/2204.06125.

[23] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10674-10685. DOI:10.1109/CVPR52688.2022.01042

[24] Esser P, Kulal S, Blattmann A, et al. Scaling rectified flow transformers for high-resolution image synthesis[C]//Proceedings of the International Conference on Machine Learning (ICML). 2024.

[25] Xu J P, Zhang X L, Li H, et al. Is everyone an artist? A study on user experience of AI-based painting system[J]. Applied Sciences, 2023,13(11):6496. DOI:10.3390/app13116496

[26] Müller V C, Bostrom N. Future progress in artificial intelligence: A survey of expert opinion[M]//Müller V.C. Fundamental Issues of Artificial Intelligence. Cham: Springer International Publishing, 2016: 555-572. DOI: 10.1007/978-3-319-26485-1_33

[27] Fjelland R. Why general artificial intelligence will not be realized[J]. Humanities and Social Sciences Communications, 2020,7(1):10. DOI:10.1057/s41599-020-0494-4

[28] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the International Conference on Machine Learning (ICML). 2021:8748-8763.

[29] Schuhmann C, Beaumont R, Vencu R, et al. LAION-5B: An open large-scale dataset for training next generation image-text models[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2022: 25278-25294.

[30] Zhang Z L, Zhao T C, Guo Y L, et al. RS5M and GeoRSCLIP: A large-scale vision- language dataset and a large vision-language model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5642123. DOI:10.1109/TGRS.2024.3449154

[31] Li J N, Li D X, Savarese S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the International Conference on Machine Learning (ICML). 2023: 19730-19742.

[32] Jia C, Yang Y F, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the International Conference on Machine Learning (ICML). 2021:4904-4916.

[33] Wang Z C, Prabha R, Huang T Y, et al. SkyScript: A large and semantically diverse vision-language dataset for remote sensing[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5805-5813. DOI:10.1609/aaai.v38i6.28393

[34] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014:740-755. DOI:10.1007/978-3-319-10602-1_48

[35] Shao S, Li Z M, Zhang T Y, et al. Objects365: A large-scale, high-quality dataset for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 8429-8438. DOI: 10.1109/ICCV.2019.00852

[36] Ding H H, Liu C, He S T, et al. MeViS: A large-scale benchmark for video segmentation with motion expressions[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 2694-2703. DOI: 10.1109/ICCV51070.2023.00254

[37] Liu C, Ding H H, Jiang X D. GRES: Generalized referring expression segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023:23592-23601. DOI:10.1109/CVPR52729.2023.02259

[38] Yuan Z H, Mou L C, Hua Y S, et al. Referring image segmentation for remote sensing data[C]//IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024:946-949. DOI:10.1109/IGA

RSS53475.2024.10642726.

[39] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2014: 2672-2680.

[40] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65. DOI: 10.1109/MSP.2017.2765202

[41] 林懿伦,戴星原,李力,等.人工智能研究的新前线:生成式对抗网络[J].自动化学报,2018,44(5):775-792. [Lin Y L, Dai X Y, Li L, et al. The new frontier of AI research: Generative adversarial networks[J]. Acta Automatica Sinica, 2018,44(5):775-792.] DOI:10.16383/j.aas.2018.y000002

[42] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2016:2234-2242.

[43] Denton E, Chintala S, Szlam A, et al. Deep generative image models using a laplacian pyramid of adversarial networks[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2015:1486-1494.

[44] Karras T, Aila T M, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2018.

[45] 胡铭菲,左信,刘建伟.深度生成模型综述[J].自动化学报,2022,48(1):40-74. [Hu M F, Zuo X, Liu J W. Survey on deep generative model[J]. Acta Automatica Sinica, 2022,48(1):40-74.] DOI:10.16383/j.aas.c190866.

[46] 谢天圻,吴媛媛,敬超,等.GAN模型生成图像检测方法综述[J].计算机工程与应用,2024,60(22):74-86. [Xie T Q, Wu Y Y, Jing C, et al. Survey of image detection methods generated by GAN models[J]. Computer Engineering and Applications, 2024, 60(22): 74-86.] DOI: 10.3778/j.issn.1002-8331.2405-0346

[47] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein GANs[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2017:5769-5779.

[48] Radford A, Metz L, Chintala S, et al. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. 2015: 1511.06434. https://arxiv.org/abs/1511.06434v2.

[49] Reed S, Akata Z, Yan X C, et al. Generative adversarial text to image synthesis[C]//Proceedings of the International Conference on Machine Learning (ICML). 2016: 1060-1069.

[50] Zhang H, Xu T, Li H S, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 5908-5916. DOI: 10.1109/ICCV.2017.629

[51] Sohn K, Yan X, Lee H. Learning structured output representation using deep conditional generative models[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2015:3483-3491.

[52] Walker J, Doersch C, Gupta A, et al. An uncertain future: Forecasting from static images using variational autoencoders [C]//Proceedings of the European Conference on Computer Vision (ECCV). 2016: 835-851. DOI: 10.1007/978-3-319-46478-7_51

[53] Theis L, Oord Aaron V D, Bethge M. A note on the evaluation of generative models[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2016:1-10.

[54] Oord Aaron V D, Vinyals O, Kavukcuoglu K. Neural discrete representation learning[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2017:6309-6318.

[55] Razavi A, Oord Aaron V D, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2019:14866-14876.

[56] Ding M, Yang Z, Hong W, et al. CogView: mastering text-to-image generation via transformers[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2021:19822-19835.

[57] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML). 2015:2256-2265.

[58] Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2019:11918-11930.

[59] Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2020.

[60] Song J, Meng C, Ermon S. Denoising diffusion implicit models[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2021:1-20.

[61] Lu C, Zhou Y, Bao F, et al. DPM-solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2024:5775-5787.

[62] Karras T, Aittala M, Laine S, et al. Elucidating the design space of diffusion-based generative models[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2024:26565-26577.

[63] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models[C]//Proceedings of the International Conference on Machine Learning (ICML). 2021: 8162-8171.

[64] Bao F, Li C, Zhu J, et al. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2022:1-39.

[65] Dhariwal P, Nichol A Q. Diffusion models beat GANs on image synthesis[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2021: 8780-8794.

[66] Zhang L M, Rao A Y, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023:3813-3824. DOI:10.1109/ICCV51070.2023.00355

[67] Tang D, Cao X, Hou X, et al. CRS-Diff: Controllable remote sensing image generation with diffusion model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024,62:1-14. DOI:10.1109/TGRS.2024.3453414

[68] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). 2017:6629-6640.

[69] Hessel J, Holtzman A, Forbes M, et al. CLIPScore: A reference-free evaluation metric for image captioning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 7514-7528. DOI:10.18653/v1/2021.emnlp-main.595

[70] Xu T, Zhang P C, Huang Q Y, et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018:1316-1324. DOI:10.1109/CVPR.2018.00143

[71] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding [C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2022:36479-36494.

[72] Yu J, Xu Y, Koh J Y, et al. Scaling autoregressive models for content-rich text-to-image generation[J]. Transactions on Machine Learning Research, 2022:2835-8856

[73] Petsiuk V, Siemenn A E, Surbehera S, et al. Human evaluation of text-to-image models on a multi-task benchmark[EB/OL]. 2022: 2211.12112. https://arxiv. org/abs/2211.12112.

本文图文责任编辑：蒋树芳 黄光玉