PlantoGraphy: Incorporating Iterative Design Process into Generative Artificial Intelligence for Landscape Rendering

RONG HUANG, The Hong Kong University of Science and Technology (Guangzhou), China

HAI-CHUAN LIN, The Hong Kong University of Science and Technology (Guangzhou), China

CHUANZHANG CHEN, The Hong Kong University of Science and Technology (Guangzhou), China

KANG ZHANG, The Hong Kong University of Science and Technology (Guangzhou), China and The Hong Kong University of Science and Technology, China

WEI ZENG*, The Hong Kong University of Science and Technology (Guangzhou), China and The Hong Kong University of Science and Technology, China

Landscape renderings are realistic images of landscape sites, allowing stakeholders to perceive better and evaluate design ideas. While recent advances in Generative Artificial Intelligence (GAI) enable automated generation of landscape renderings, the end-to-end methods are not compatible with common design processes, leading to insufficient alignment with design idealizations and limited cohesion of iterative landscape design. Informed by a formative study for comprehending design requirements, we present *PlantoGraphy*, an iterative design system that allows for interactive configuration of GAI models to accommodate human-centered design practice. A two-stage pipeline is incorporated: first, *concretization* module transforms conceptual ideas into concrete scene layouts with a domain-oriented large language model; and second, *illustration* module converts scene layouts into realistic landscape renderings using a fine-tuned low-rank adaptation diffusion model. *PlantoGraphy* has undergone a series of performance evaluations and user studies, demonstrating its effectiveness in landscape rendering generation and the high recognition of its interactive functionality.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; **Interaction design**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Landscape rendering, large language model, scene graph, generative artificial intelligence

1 INTRODUCTION

While Goethe refers to architecture as frozen music, Filor draws an analogy between landscape design and ballet, which aims at discovering how the elements of nature can be recombined responsive to both planned and unforeseen uses on a daily and seasonal timescale [14]. The traditional landscape design process relies on the creativity and expertise of designers to create useful, comfortable and attractive spaces [3], which can be divided into four main stages: a) initial conceptualization, b) design development, c) 3D modeling, d) rendering, as demonstrated in Fig. 1(bottom). Specifically, designers develop a preliminary concept taking into account site character, user acquirement and design vision, then refine the design concept into a detailed plan that includes specific design elements, such as plant selection and configuration. The process outputs human-perspective landscape renderings that depict realistic images of a planned site with all plants, allowing stakeholders to evaluate the visual quality of landscape designs.

---

*Wei Zeng is the corresponding author.

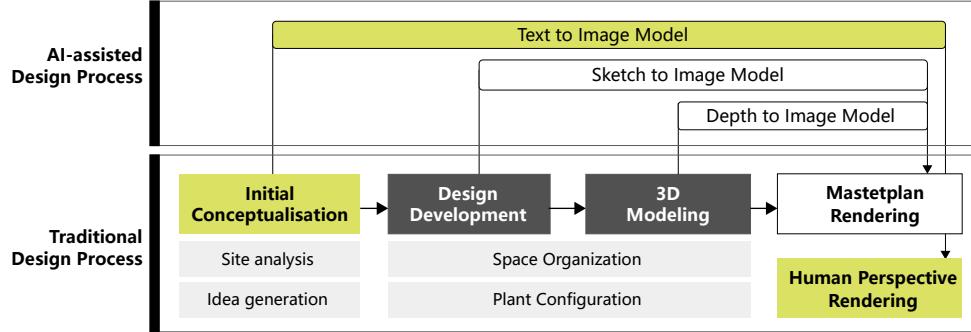Rong Huang, Hai-Chuan Lin, Chuanzhang Chen, Kang Zhang, and Wei Zeng

Fig. 1. Comparison of traditional and AI-aided design process for landscape rendering. Various AIGC models can be applied to different stages, yet these methods operate in an end-to-end manner without integrating iterative design.

With the advances in generative artificial intelligence (GAI), AI-based image generation tools have emerged to facilitate the landscape design process. Pre-trained text-to-image generation models, such as Stable Diffusion [37] and DALL-E-2 [36], can take descriptions for a garden like "*a realistic picture of a landscape design with trees, including a dogwood with pink flowers, flowering plants such as white tulips and daisies*", and produce corresponding landscape rendering. Due to the flexibility, GAI tools can play varying roles at different stages of the design process, as illustrated in Fig. 1(top). Within the context of human-perspective landscape rendering, AI is acting as co-creators with the designers, refining the abstract concept and generating human perspective renderings through text-to-image models [36, 37].

However, existing GAI tools primarily work in an end-to-end manner, lacking the flexibility necessary for designers to incorporate common design practices that iteratively refine the outputs. In particular, we identify the following limitations by existing AI-supported designing process from literature review (Sect. 2) and a formative study with landscape designers (Sect. 3). 1) *Absence of interactive functionality for iterative design.* Landscape design is an iterative process shaped by both concept and form, aiming at finding the approaches that best respond to expectations [14]. Existing end-to-end approaches focus on design control within text editing, along with in-coherent results in multiple generations, causing absence of interactive functionality for iterative design. 2) *Insufficient sensitivity to design arrangement of elements.* Landscape designers are tasked with selecting plant species and determining the optimal combination expressed using directional words and commonly used plant words, which may not be accurately interpreted by text-to-image models. For example, existing AIGC tools are not compatible with descriptions like "*the daisy is located below the dogwood, and the white tulip is positioned to the right of the daisy.*"

To address these requirements from landscape designers, we present *PlantoGraphy*, an intelligent system with interactive functions for iterative refinement and an enhanced comprehension of landscape scenario descriptions. As illustrated in Fig. 2, *PlantoGraphy* has two main modules: 1) *Concretization module* (Sect. 4.3) that transforms idea descriptions to scenario layout. The module introduces the concept of the scene graph, which represents landscape design as a structured, semantic description of the objects and their relationships within a scene. Employing a dataset comprising scene descriptions, scene graphs, and layouts, this module utilizes a large language model (LLM) to transform the designers' descriptions into layouts with the guidance of scene graphs, which provide more interpretable and controllable information about the landscape scene for guiding the rendering generation task. 2) *Illustration module* (Sect. 4.4) that generates layout-guided landscape rendering. We harvest a vegetation dataset commonly used in landscape design, with the participation of expert designers. Utilizing this dataset, the latent diffusion model [28, 37] is fine-tuned through a Low-rank Adaptation (LoRA) model [20], which enhances the model's ability to generate specific
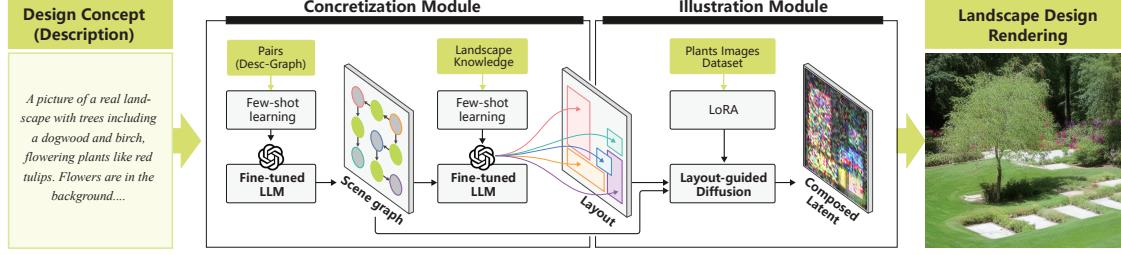
Fig. 2. Overview of the system workflow. *PlantoGraphy* incorporates a two-stage framework to transform design concepts described in textual content into realistic images of landscape rendering. First, the *concentization* module leverages a graph-enhanced LLM to transform design descriptions into layouts, using scene graphs to improve the comprehension of user input. Next, the *illustration* module employs a fine-tuned LoRA model to generate realistic landscape renderings based on the layout.

plants within the landscape scenario with high precision and accuracy. *PlantoGraphy* also incorporates a web-based user interface (Sect. 4.2) that enables users to easily generate landscape design renderings and iteratively refine them by interacting with the graph and layout components.

A thorough assessment of *PlantoGraphy* has been conducted from various perspectives. For concretization module assessment (Sect. 5.1), two subjective experiments with quantitative metrics pertaining to layout were conducted, highlighting the essential role of incoperating scene graph in layout prediction and injecting domain knowledge using reasoning methods. For illustration module evaluation (Sect. 5.2), three experiments in both objective and subjective perspectives were carried out, with results demonstrating the effectiveness of *PlantoGraphy* in improving coherence in the results of multiple generations and comprehending designers' intents. To further evaluate the proposed AI-assisted landscape design process, we conducted a within-subjects study (Sect. 6) that involves creating landscape renderings using *PlantoGraphy* and traditional design software separately. The results suggest that *PlantoGraphy* holds varying degrees of value at different design stages. The findings also highlight human-centered AI assistance in design, emphasizing the importance of supporting iterative refinement of designs at each stage of the design process.

The major contributions and novel aspects of this work include:

- **Framework**. We propose a novel framework that incorporates an iterative design process into the AI-assisted designing pipeline for landscape rendering. This framework maximizes control over the coherence in multiple generations of results, and addresses the challenge of the generative model's insensitivity to commonly used orientation words and plant terms in landscape design scene descriptions.

- **System**. We introduce an interactive system, namely *PlantoGraphy*, to support the landscape rendering generation and iterative refinement. The system is built upon a custom dataset created in collaboration with expert designers that comprises realistic plants rendering images. *PlantoGraphy* speeds up the iteration by allowing designers to set constraints through interaction, promoting the output design to be more consistent with their expectations and increasing the involvement of landscape designers in the AI-assisted designing process.

2 RELATED WORK

2.1 Generative AI for Creative Design

Generative models are capable of generating new data points that are similar to the training data set [15]. The recent surge in GAI has ignited growing interest across various creative design fields, including fashion design [10, 50], UI design [24], and visualization design [52]. GAI serves to either automate specific tasks or to foster the exploration

of creative ideas [26] during the creation phase among various design processes [16, 41]. Recently, there has been a growing interest in diffusion models [18, 42] among designers and artists. These include text-to-image (T2I) diffusion models [33, 36, 37] that leverage textual prompts to direct the image generation process, and other conditional generative models like ControlNet [56], T2I adapter [32] and GLIGEN [28], which allow more fine-grained input conditions. While the remarkable efficiency and generation ability have been shown in technique manner, some limitations impede the usability of GAI in practical way of design. The first and foremost reason is that, most diffusion models are realized in an end-to-end manner with input text prompts and output images, which overlook the "exploration" stage as one of the most important parts in creative process [21]. In addition, the lack of controllability cannot support stylistic consistency outputs, which is crucial for iterative design in practical scenarios [43].

This study aims to develop an AI-assisted design system for landscape rendering. Three specific challenges are addressed: integrating the design process into the model, obtaining suitable training datasets for this specific context, and exerting control over factors such as layout, plant positioning, and sizing. While LLM-grounded diffusion [29], along with some other models [13, 59], claim to offer control over the shape, position, and appearance of generated objects, our experiments have demonstrated that their outcomes are hard to meet expectations of domain experts. Furthermore, these methods do not encompass the final stage of rendering realistic landscape views. To bridge this gap, we contribute a LoRA finetuning model trained on a landscape dataset that we have curated ourselves.

2.2 Reasoning in Large Language Model

LLMs have demonstrated promising results across a range of downstream tasks [38]. The capabilities have been exemplified by the GPT series [4, 34], revealing emergent abilities that become apparent as the scale of training reaches a certain threshold [47]. While LLMs are originally designed for text-based tasks, researchers have also ventured into other domains like vision-language reasoning. For example, visual GPT [49] fuses ChatGPT with visual foundation models to tackle vision-language tasks in an interactive manner. Nevertheless, one significant challenge impeding the practical application of LLMs is their limited ability for reasoning. In-context learning (ICL) [12] addresses this challenge by enabling LLMs to generate expected outputs when given input text. Studies have showcased the effectiveness of LLMs in solving complex reasoning problems through ICL [45]. Chain-of-thought (CoT) [48] approaches are introduced to bolster the reasoning capabilities in complex tasks by introducing intermediate reasoning steps that lead to the final output. CoT can be used with ICL in two main ways: Few-shot CoT and Zero-shot CoT [27]. Few-shot CoT applies the step-by-step reasoning in the form of $< input, output > \rightarrow < input, CoT, output >$, whilst Zero-shot CoT directly generates intermediate reasoning steps to derive the answers, as exemplified by the phrase *"let's think step by step"*. Building upon the CoT framework, several works such as Auto-CoT [57], Tree-of-Thought [55], multimodal-CoT [58], and multilingual-CoT [39] have been proposed to elicit reasoning abilities of LLMs in various tasks.

LLM-grounded diffusion [29] combines GPT and diffusion models to perform conditioned text-to-image tasks, which can be applied to landscape rendering tasks. However, this model does not inherently support iterative design process that is required by landscape designers. The central challenge revolves around finding an effective method to integrate design rationales in the design process, into LLM-based design. To address this limitation, we leverage the scene graph concept, to represent plants within a landscape as entities and their spatial relationships as edges. Experimental results demonstrate that this innovative approach can produce precise and realistic landscape renderings based on designers' intentions, and more importantly, facilitate iterative design process.

2.3 Scene Graph

Scene graph is a structured data model employed to represent spatial relationships and semantic information of objects within a scene [22, 23]. The ability to intuitively depict spatial relations has made scene graphs an emerging topic in both computer vision and AI research [6]. As a graph-based representation, scene graphs can be seamlessly integrated with deep-learning-based generative models. With advancements in conditional image synthesis, techniques have evolved to render images by conditioning generative models on scene graphs, such as GANs [2, 44] and diffusion models [53]. These methods allow designers and artists to work in a more abstract and intuitive manner by manipulating objects and their relationships rather than directly editing pixels or vertices.

Notably, the concept of the scene graph is highly aligned with the abstract bubble maps commonly employed in design sketch [11, 17]. The integration of scene graphs with generative models has the potential to enhance reasoning capabilities and provide users with greater control. Scene graphs have been effectively employed in image generation through the prediction of bounding boxes and segmentation masks [2, 22]. However, these methods primarily rely on GAN-based approaches, which can present challenges in achieving high-resolution results and recognition of the rendering orders. We address these challenges with LoRA diffusion models for high resolution renderings and an instance-based latent composition for controlling the overlapping order of objects. Conversely, this study seeks to explore the possibility of leveraging scene graphs in conjunction with LLMs and diffusion models. This represents a new and inherently challenging avenue of research, particularly for creative design.

3 DESIGN STUDY

This section presents a group interview with landscape design experts aimed at gaining insights into the design prerequisites for collaborative efforts between humans and AI in landscape rendering (Sect. 3.1). In light of the findings, we consolidate design goals to be achieved for *PlantoGraphy* (Sect. 3.2).

3.1 Group Interview

We conducted online interviews with five landscape design professionals (2 designers in employment *U1* and *U2*, 2 landscape design students *U3* and *U4*, 1 landscape design researcher *U5*). All designers have more than three years of landscape design experience. Each interview lasted 40-60 minutes. Specifically, as participants *U2* and *U4* had no prior experience using AI tools, we provided an introduction to existing tools and basic instructions before interviews. We designed a question outline mainly focusing on three topics based on the relevant literature review and asked each designer to answer questions according to their design experience with no specific limitation: 1) the general workflow of landscape design, 2) collaboration with artificial intelligence, and 3) the pros and cons of existing AI-aided landscape designing tools according to their using experience.

At the end of the interview, we summarized the core insights based on the designers' feedback as follows:

Landscape designing process. We first surveyed relevant literature and industrial standard documentation to summarize the traditional landscape design process and what contents are usually included. With details supplemented by the designers' practice experience, the traditional landscape designing process mainly comprises four stages (Fig. 1(bottom)): First, designers determine the initial design concepts according to the information collected from site analysis and cultural study. Second, they develop designs by organizing space and flow on the sketch of masterplan. Third, designs will be modeled in professional 3D modeling software and refine the construction detail. Finally, designers

will render the 3D model into a real scene through rendering engines to test the visual experience from a human perspective. It's worth noting that participants mentioned that in the practice project,

"...design requirements are always adjusted, the aforementioned process is cyclical rather than sequential, until an optimal fit between concept and approach is achieved, as refereed iterative design." - U1 & U2

Collaboration with AI. AI can be utilized to assist at different stages of landscape design, as illustrated in Fig. 1. When the input is a scenario sketch or a depth map, the design is completed by designers while the AI serves as an acceleration tool in the drawing process. Another situation is that the AI acts as a co-creator for the landscape designer when the input is an initial conceptual description expressed in textual prompts. Participants *U3* and *U4* expressed a preference for this co-creative approach, which allows designers to impart their ideas by delivering diverse interpretations of the textual concepts. This aligns with common design practices where designers often receive vague and abstract descriptions from clients. However, *U1* and *U2* expressed concerns about the potential loss of control for designers when using graphical tools to bypass essential steps in the landscape design process. They all agree that when designers are deprived of multi-dimensional control over the outcome, the meaning of design is lost.

"I feel like I'm more of a user than a designer." - U1

Existing tools for landscape design. All participants agreed that current tools can provide designers with multiple solutions to inspire them, but they rely heavily on text input, which is not always accurately understood by the model, particularly with regard to orientation descriptions. Besides, participants expressed a preference for more interactive functions of graphical interfaces complemented with textual prompts that convey the logic of designers' thinking, allowing designers to express their design intent in multiple ways (*U3*). *U1* and *U2* noted that each time they refined the description, the outputs by existing models varied drastically, far away from designers' needs to iteratively refine the design to accommodate changing conditions. Additionally, designers need to select plant species considering the environmental conditions and characteristics of species and determining the best location for each plant based on its environmental requirements and aesthetic considerations.

"Current model has limited plant species available resulting in incorrect plant species generation." - U1 & U3 & U4

3.2 Design Goals

Our objective is to harness human-AI collaborations to support the creative design process for landscape renderings. This process entails progressing from general concepts to detailed elements, with iterative adjustments and refinements made along the way [41]. In an AI-assisted design approach, the goal is to minimize repetitive tasks while enhancing the designer's creativity, rather than replacing them [1, 5]. To achieve this, the system needs to enhance its ability to interact with designers, accepting their guidance and actively suggesting possible solutions, thus completing a feedback loop in the creation process [40, 41, 54]. In line with the HCI community's exploration of augmenting user control over models, enhancing designer involvement in the AI-supported creating stage can fully leverage specialized knowledge [40]. The graphical user interface (GUI) plays a crucial role as a user-friendly and efficient interaction approach to accept inputs from designers. This includes using buttons and sliders for inputting basic settings [9], text editors for expressing design concepts [7, 51], and graphical control for modification [8, 46].

Based on the feedback gathered from the design study and existing design practices, we formulate a set of design goals for our intelligent landscape design system.

- **G1: Improvement of the result's graphical coherence to support iterative design.** To meet the iterative modification needs of designers, our system should enhance the graphical consistency of multiple generating

results with the exception of the modified portion when user input is micro-changed, *e.g.*, slightly changing the plants' position and sizes.

- **G2: Enhanced model's comprehension of landscape design descriptions.** Our system should be equipped with a model capable of comprehending design descriptions in landscape scenarios, particularly orientation words that are commonly used by designers. Additionally, it should support accurate generation for expert commonly-used plant species.
- **G3: Interactive editing function for in-depth designer participation.** Our system should offer multiple ways for landscape designers to express their design requirements. In line with the thinking logic of designers, enabling textual prompts as input is necessary. A graphic-based editing function is a suitable addition, enabling designers to interact with the system in a more intuitive and visual manner.

4 PLANTOGRAPHY SYSTEM

This section introduces *PlantoGraphy*, a novel system that supports the steerable generation of landscape rendering from scene description and facilitates iterative design by interactive functionality. We summarize how the system design responds to the design goals in Sect. 4.1, followed by a detailed description of the interface design (Sect. 4.2). Then we introduce the main modules of the system: a *concretization* module (Sect. 4.3) and an *illustration* module (Sect. 4.4).

4.1 System Overview

PlantoGraphy is crafted for experienced landscape designers proficient in generating landscape renderings using the system and iteratively refining the results according to design requirements through the interactive editing module. To accomplish the goal, we design a two-stage approach as demonstrated in Fig. 2:

- **Concretization module.** The module harnesses domain-specific LLMs to translate users' conceptual ideas, as depicted in textual descriptions, into concrete scene layouts. An innovation introduced here is the incorporation of *scene graph* as an intermediary link between text and layouts. Scene graph representation enhances the LLM's ability to comprehend landscape design descriptions (**G2**) when compared to direct conversion. Furthermore, the scene graph acts as a medium for users to customize their designs, complementing text-only user interactions to improve graphical coherence (**G1**) and enhance interactive editing (**G3**). The module offers *scene graph* and *layout* visual representations, to facilitate the iterative idea expression in an intuitive and interactive manner. In the backend, two LLM-powered generators guided by well-structured prompt templates support the transformation between text, scene graph, and layout. These prompts incorporate both reasoning enhancement techniques and domain-specific knowledge relevant to landscape design. Furthermore, interactive features enable designers to make adjustments intuitively by editing the graph and layout. This includes actions such as adding or removing nodes and resizing elements to accommodate evolving design needs.
- **Illustration module.** The module employs a LoRA finetuning model to convert scene layouts into realistic landscape renderings. One of the significant challenges faced in this process is the scarcity of available training samples. To address this challenge, we have curated a vegetation dataset commonly utilized in landscape design with the invaluable input of expert designers. The dataset encompasses a wide range of plant types, along with detailed attributes like common plant composition patterns. This makes the generated outputs exhibit greater graphical coherence with user descriptions, aligning with **G1**. Furthermore, we tackle the issue of overlapping object order, by implementing an instance-based latent composition process. This module offers full automation,

thereby saving designers time in design development and 3D modeling and empowering them to visualize how their conceptual ideas translate into human-perspective views.
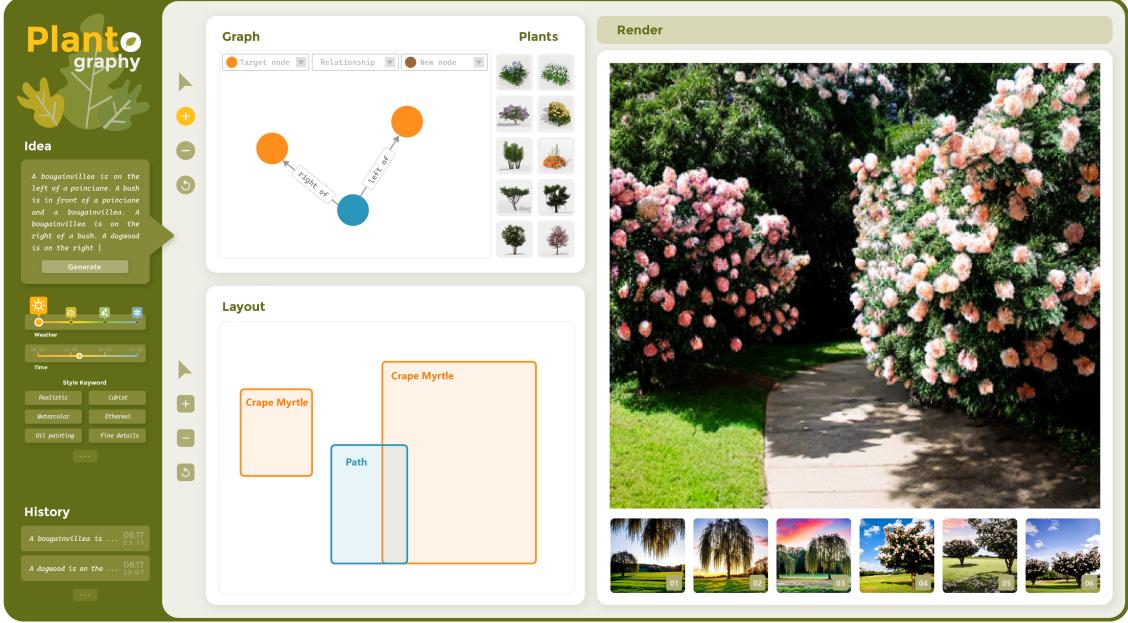

Fig. 3. Interactive visual interface for *PlantoGraphy*. Users can input textual descriptions of the scene in the *Text Panel* and customize the design by manipulating the scene graph in the *Graph Panel* and updating the layout in the *Layout Panel*. The rendering results are presented in the *Rendering Panel*.

4.2 User Interface

An interactive interface is strategically crafted to bridge the gap between the rapid generation capabilities of current generative models and the controlled, iterative nature of traditional design processes. It empowers professional designers to harness the time-saving advantages of text-to-image technology while retaining as much control over the design development as traditional methods allow, ensuring a workflow that is both efficient and aligned with the natural progression of idea development. The interface comprises four panels depicted in Fig. 3, allowing user to develop their idea in a procedural, controllable and interactive way: text, graph, layout, and rendering.

- **Text Panel.** The Text Panel initiates the design process. Designers can enter descriptive text, including plant species, quantities, and positional relationships between objects in the desired scene. For example, *"A realistic picture of a landscape design with trees, including a dogwood with pink flowers, flowering plants such as white tulips and daisies. The daisy is located below the dogwood, and the white tulip is positioned to the right of the daisy."* To assist designers in avoiding the need to repeatedly input common constrain prompts, we offer commonly used options related to landscape design, including sliders for time and season, as well as rendering styles. Additionally, at the bottom of the panel, we provide a historical module to store previous design flows, allowing designers to iterate designs across different times and projects

- **Graph Panel.** The Graph Panel organizes and visualizes the scene graph generated from designers' descriptions, while also enabling iterative adjustments to the scene graph. Each individual plant is represented by a node on
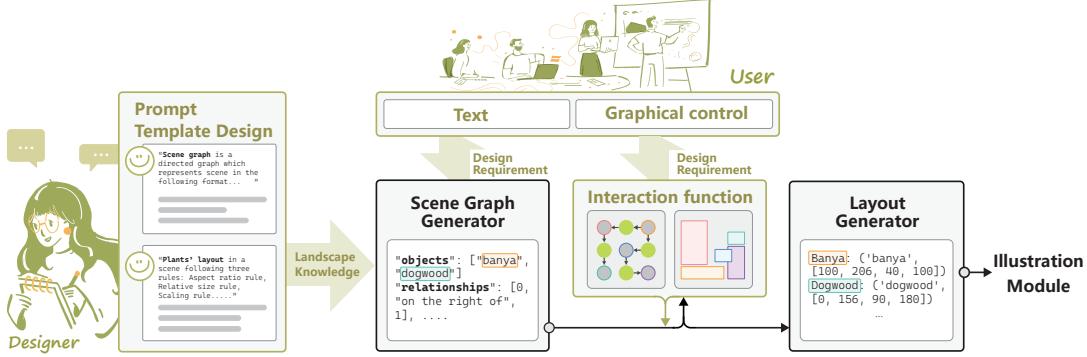
Fig. 4. Framework of idea concretization via expert-engaged interaction with LLMs for scene graph generator and layout generator.

the graph, and the oriented edge represents the relationship between nodes. The label on the edge is a preposition indicating the orientation relationship between nodes. Designers can create and manipulate nodes and edges to represent plants and their relationships. A plant database panel is provided on the right side, allowing designers to directly drag the plant image as input.

- **Layout Panel.** The Layout Panel presents the inferred layout of the scene based on textual inputs and the scene graph, finalizing spatial details before transitioning to concrete landscape renderings. The panel displays each node in the graph as a box, with absolute positional coordinates and accurate plant size, inferred from the concretization module. Designers can adjust positions and sizes of elements to finalize the spatial layout.

- **Rendering Panel.** The Rendering Panel showcases the final landscape renderings generated by the system based on the intermediate scene graph and layout. The generated images by the iterative modifications are also stored at the history view in the bottom, which can be easily clicked on by the user to make comparisons.

4.3 Idea Concretization via Large Language Model

The *concretization* module is designed to transform textual scene descriptions into concrete layouts of design element bounding boxes, serving as visual guidance and conditional input for the generative model in subsequent sections. Drawing insights from the group interview (Sect. 3.1), where designers expressed the expectations of supporting iterative design and improving the comprehension of design intentions, we propose to integrate scene graph as an intermediate component in the process. The scene graph aids in organizing design elements at an abstract level, avoiding early constraints imposed by concrete details and seamlessly bridging the transition from textual conceptual ideas to final layouts. As such, we develop two LLM generators: the first one transforms scene descriptions to a scene graph, and the second one transforms the scene graph to a layout separately, through prompt engineering. Specifically, the second generator is reasoning-intensive, requiring spatial and domain-specific reasoning in landscape design. The LLM needs to infer the relative positions of plants in a two-dimensional space, deduce the size of each plant based on landscape knowledge, and refine the layout considering various constraints. The overall framework is illustrated in Fig. 4.

Consequently, we have developed two distinct prompt templates, as shown in Fig. 5. Both templates comprise four main components: the task description, constraints, contextual information, and demonstrations. To equip the LLM's with domain-specific reasoning capabilities, the prompt for the graph-to-layout generator includes a landscape knowledge component as contextual information. Below, we use the graph-to-layout generator as an example to illustrate each component. The complete prompt templates are shown in the Supplementary Material.

Task Description

a Prefix

You are an expert in bounding box generation.

b General description

Giving you a scene graph and a set of generation rules, your task is to generate bounding box for each object in the given scene graph based on the spatial relation descripted in scene graph and generation rules.

Each bounding box should be in the format of (object name, [top –left x coordinate, top –left y coordinate, box width, box height]) and include exactly one object.

c Step by step instruction

Your task should be done in three steps.

Step 1: Determine the position of each object based on scene graph. From left to right, the x position decreases.

Step 2: Determine the size of each object based on the aspect ratio rule and relative size rule.

Step 3: Adjust the size of each object based on the scaling rule. (determine the distance relationship between objects, then scale the object behind)

d Constraints

The images are of size 512x512, and the bounding boxes should not overlap or go beyond the image boundaries.

e Contextual information

Aspect ratio rule:
1. Banya have aspect ratio as 4:10
2. Dogwood have aspect ratio as 2:4
    ...

Relative size rule:
1. The height of dogwood is twice as high as banya.
    ...

Scaling rule:
If there is relation that present the front and back relation, such as [a, in front of, b], the size between a and b should have a scaling relation. The object behind should scale the width and height smaller than before, keeping the aspect ratio.   ...

f Demonstration

Input scene graph: "objects": ["banya", "dogwood"]
                                      "relationships": [0, "on the right of", 1], [0, "in front of", 1]
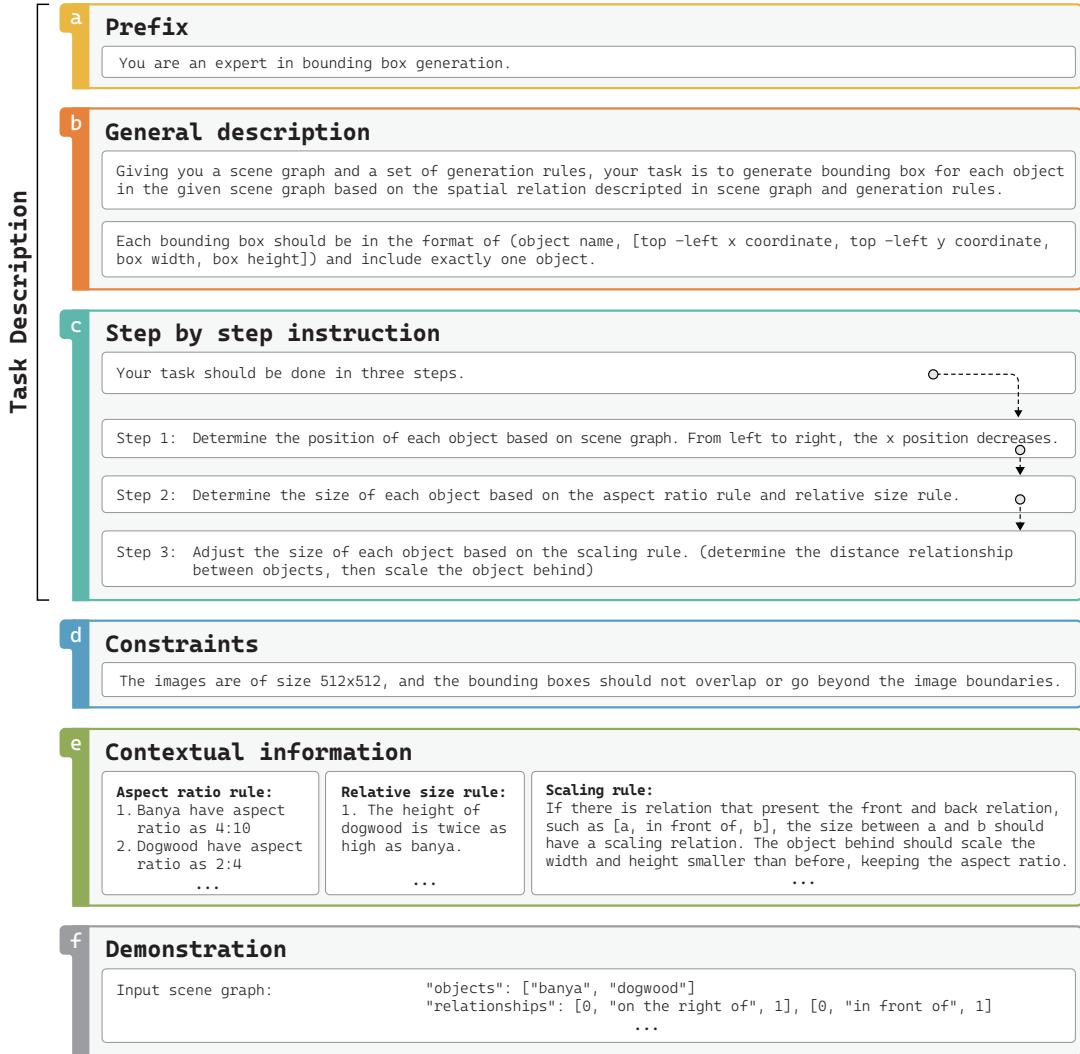   ...

Fig. 5. Prompt template for landscape scene layout generation. The template consists of four main components: task description, constraints, contextual information, and demonstrations.

- *Task Description:* The prompt commences with a prefix prompt (Fig. 5 (a)) followed by a general description (Fig. 5 (b)). The prefix prompt improves LLM's ability to perform specialized tasks with role conditioning. The general description elucidates the input, output, and output formats. We transform non-natural language entities, such as scene graphs and layouts, into a sequential format to enhance comprehensibility for the LLM. The scene graph structure is linearized as a series of triples $< a, relation, b >$, where a and b represent nodes in the graph, and $relation$ corresponds to the edge connecting a and b. The layout is represented as $[objectname, [x, y, width, height]]$. To bolster reasoning capabilities during the generation process, we employ a chain-of-thought strategy as a step-by-step instruction (Fig. 5 (c)).
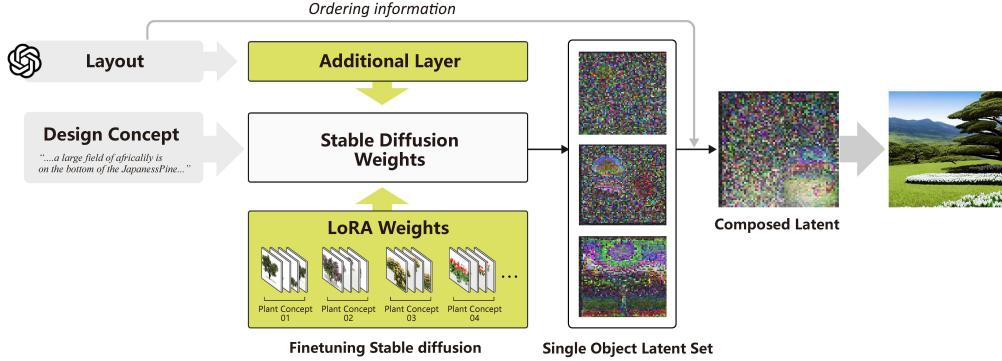
Fig. 6. Framework of landscape illustration via layout-guided landscape rendering generation.

- *Constraints:* The constraints (Fig. 5 (d)) define the boundaries within which the LLM is authorized to perform reasoning. Our prompts incorporate various types of constraints, such as restrictions in vision tasks and limitations on the number of generated elements in graphs.
- *Contextual information:* While LLMs possess reasoning abilities, directly expecting them to reason with unfamiliar professional knowledge absent in the pre-trained model can be challenging. Therefore, we introduce three key rules in landscape render generation for landscape design as contextual information for the LLM to reference during reasoning. These rules encompass the aspect ratio of each plant, the relative sizes between plants, and the scaling effect for perspective relationship (Fig. 5 (e)).
- *Demonstration:* We provide 5 question-answer examples to facilitate the LLM's reasoning ability and ensure alignment with the desired answering format through few-shot learning (Fig. 5 (f)).

4.4 Landscape Illustration via Customized Diffusion Model

The *illustration* module takes the layout combined with textual prompts as inputs to produce realistic landscape renderings. The layout dictates the placement of plants, while the prompt specifies overall scene conditions like season and weather. To address the limitations of reliably generating specific plants and accurately deducing overlapping order faced by existing generative models, we have developed a more controllable framework, as illustrated in Fig. 6. This framework is composed of several components. First, we employ GLIGEN [28] to act as the foundational model for open-set, layout-guided image generation. We train LoRA models [20] to fine-tune GLIGEN, enhancing its ability to generate domain-specific plants. Additionally, an instance-based latent composition process is employed to manage the depth relationships among objects. This integrated approach enhances the model's capacity to produce controllable and consistent landscape renderings, tailored to the requirements of professional landscape design.

4.4.1 Layout-guided diffusion model. GLIGEN is used as its base model for high-quality, layout-guided image generation. GLIGEN enhances pre-trained models by integrating location inputs through a new trainable gated-attention layer, preserving the original pre-trained weights. This enhancement allows the framework to effectively generate images based on both textual descriptions and layout information, while maintaining robust zero-shot capabilities.

4.4.2 Model fine-tuning for better plant generation control. To improve plant generation, we further perform fine-tuning of the base model using a self-curated dataset with the help of landscape designers, who provide their expertise and creativity in selecting the best plants and arrangements for each scene. The dataset comprises hundreds of realistic rendering images in various scenes and includes 11 types of trees and 11 types of bushes, with each plant type consisting

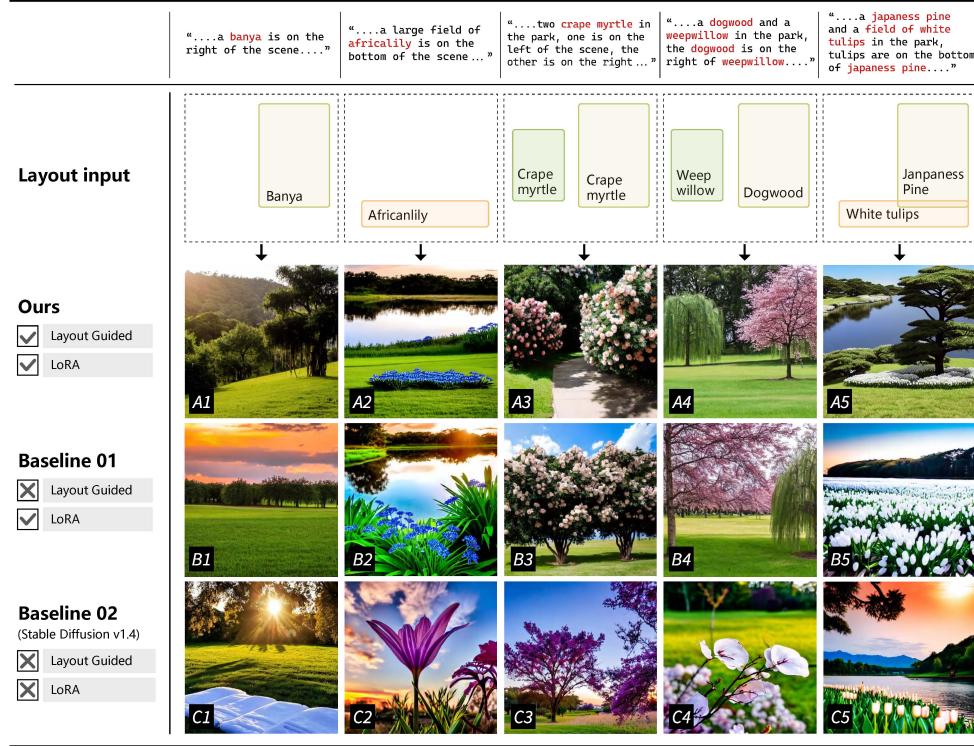Fig. 7. Results of with (bottom) and without (top) instance-based latent composition.

Fig. 8. Comparison of generated results between our method to existing approaches.

of about 20 images. The dataset is used to fine-tune LoRA, an effective fine-tuning method that updates only a small set of parameters in large pre-trained models. We train LoRA models based on Stable Diffusion UNet and merge the LoRA weights with the frozen Stable Diffusion UNet part of GLIGEN. Experiments show that the new types of plants introduced by our LoRA model can be correctly generated with layout conditions, even for never encountered concepts.

4.4.3 Instance-based latent composition. Drawing inspiration from [29], we adopt an instance-based latent composition method, to address the limitation of GLIGEN's insensitivity to the front and back positions. The method treats each plant as a separate foreground instance, distinct from the background. The generation process is outlined as follows:

(1) Utilize GLIGEN to generate images for each plant in the corresponding bounding box separately.
(2) Employ SAM [25] to segment the largest object inside the bounding box for each plant. Typically, the largest object corresponds to the target plant due to the characteristics of GLIGEN. Store the segmented object masks.
(3) Utilize DDIM inversion to retrieve the latent representation of each plant image.
(4) Determine the order of the plant objects within the scene graph, prioritizing them from the back to the front.
(5) Extract the latent within the segmented object mask for each latent representation of the plant image. Sequentially replace the extracted latent to the corresponding area (e.g., the segmented object mask) with a random Gaussian noise. The latent of the foremost plant object will be the last one to overlap. After this replacing process, the resulting latent representation is referred to as composed latent.
(6) Utilize GLIGEN once again to generate the landscape with the composed latent as initial latent. During the reference, all the areas containing plants will be frozen for certain steps.

Fig. 7 compares the results by the models with and without instance-based latent composition, using the same prompts and seeds. The results indicate that the landscapes by ours adhere to appropriate foreground and background relations, as guided by the scene graph and the front-to-back order of the plant objects. Unlike simply overlaying objects in the RGB pixel space, the overlapping in latent space ensures greater coherence with several inference steps. Fig. 8 presents generated landscapes based on the input prompts and layout, by our method (row 3) and two baselines (rows 4 & 5). The results demonstrate the necessity of incorporating layout guidance and model fine-tuning in our framework.

5 EXPERIMENTS FOR MODULE EVALUATION

We conduct experiments to evaluate the performance of the proposed modules. 1) Concretization module evaluation (Sec. 5.1): Three-dimensional quantitative metrics are employed to assess the accuracy of the LLM-powered generator in transforming text into the scene layout. 2) Illustration module evaluation (Sec. 5.2): We utilize structural similarity index metric (SSIM) [19] and conduct a controlled within-subjects user study, to evaluate the graphical coherence of the results generated during iterative design (Sect. 5.2.1). We further conduct a user study to verify whether our method succeeds in enhancing the models' comprehension of plant-related knowledge (Sect. 5.2.2).

5.1 Experiment 1: Concretization module evaluation

Metrics. Following [29], we examine the performance of the *concretization* module for layout prediction from three dimensions: *object-attribute assignment, spatial reasoning*, and *perspective relational reasoning*, as follows:

- *Object-attribute assignment* is evaluated by two metrics: First, *correctness of aspect ratio* [60] measures the correctness of aspect ratio of a single plant, to assess object distortion in the generated renderings; and second, *correctness of relative areas* measures the correctness of relative size between plants, to assess visual coherence, balance, and logical structure of the plants. Specifically, to measure *correctness of aspect ratio*, we calculate L1 error between the aspect ratio of a ground-truth plant (denoted as AR_{gt}) and the one of the generation (denoted as AR_{gen}). A single plant aspect ratio is correct if $\|AR_{gt} - AR_{gen}\|_1 < \theta$, where θ is set to 0.05. We randomly sample 100 scene layouts as ground truths, and corresponding generation results from the generation method. A

Table 1. The quantitative comparison between ours to ablation methods.

Methods	Correctness of aspect ratio	Correctness of relative areas	Correctness of relative positions	Application of scaling rule
Zero-shot GPT 3.5 (w/ domain knowledge)	51	56	58	66
Few-shot GPT 3.5 (w/ domain knowledge)	100	100	91	92
Zero-shot GPT 4.0 (w/o domain knowledge)	0	48	79	70
Zero-shot GPT 4.0 (w/ domain knowledge)	82	94	72	70
Few-shot GPT 4.0 (w/ domain knowledge)	**100**	**100**	**92**	**93**

generated scene layout have correct aspect ratio if all plants in the generated scene layout have correct aspect ratios as the ground truths. The *correctness of relative areas* is computed similarly.

- *Spatial reasoning* refers to the comprehension of the relative positions of plants, which is evaluated by the metric *correctness of relative positions*. Specifically, we utilize six common relative positions of 'left', 'right', 'top', 'bottom', 'behind', and 'in front of'. For the metric, we use the same 100 scene layout samples and measure the relative positions of all pairs of plants based on their bounding boxes, and check if the corresponding results by the generation method have the same relative position. Specifically, if bounding boxes of two plants overlap, we manually check their relative position and categorize the relation as either 'behind' or 'in front of'.

- *Perspective relational reasoning* further considers the correctness of scaling size due to perspective effects, for which objects that are further away (from human perspective) should have smaller relative sizes. Here, we first compute the ground-truth relative size ratio $SR_{gt} = size(A_{gt})/size(B_{gt})$ between two plants A_{gt} and B_{gt} based on a pre-defined dictionary of expected sizes according to domain knowledge. Then, we compute the actual relative size ratio $SR_{gen} = size(A_{gen})/size(B_{gen})$ from the two generated plants. Finally, we check the consistency of z-order relation between plants A and B with the order relation between SR_{gt} and SR_{gen}, namely A in front of $B \Leftrightarrow SR_{gt} < SR_{ge+n}$, or A behind $B \Leftrightarrow SR_{gt} > SR_{gen}$. For the metric, we use the same 100 scene layout samples and corresponding generated results. A generated scene layout have correct perspective relation if all pairs of plants in the scene layout have correct perspective relation.

The success of scene layout generation depends on whether these aspects are correct in the generated layout. In the experiment, the higher number of successful results in each aspect, the better performance of the model.

Conditions. We examine the performance of our method in comparison in five conditions: 1) zero-shot GPT3.5 with landscape knowledge, 2) few-shot GPT3.5 with landscape knowledge, 3) zero-shot GPT4.0 with landscape knowledge, 4) zero-shot GPT4.0 without landscape knowledge, and 5) few-shot GPT4.0 with landscape knowledge. To ensure a fair comparison, each condition uses basically the same prompt architecture in Fig. 5 while only adding/removing the demonstration or background knowledge for the testing. For each condition, we generate 100 samples for measurement, based on sample prompts generated through a template that takes random plants and relations as text description.

Results. Table 1 presents the count of successful results in four aspects between different models and methods. First, it is observed that both GPT-4.0 and GPT-3.5 achieve high generation performance with only a few errors in the generated results. The gap between GPT-4.0 and GPT-3.5 is not prominent, possibly because our prompts enable the GPT models to reach a certain high level at which the differences become negligible. For the zero-shot method, there is a significant gap between GPT-4.0 and GPT-3.5, suggesting that GPT-4.0 has superior zero-shot capabilities. Focusing on each aspect, we note that the zero-shot GPT-4.0, without landscape knowledge for reference, generates zero correct aspect ratios and has a low correctness rate for relative areas when relying solely on common sense. This highlights the effectiveness of injecting domain knowledge into our method. On the other hand, for position reasoning and perspective reasoning, which involve common sense reasoning, the zero-shot GPT-4.0 outperforms the zero-shot GPT-3.5. It is evident that the

Table 2. Mean SSIM of images generated by two models.

	C1	C2	C3	C4	C5	All
Baseline	0.0956	0.1017	0.1123	0.1138	0.1387	0.1091
Ours	**0.1201**	**0.1466**	**0.1858**	**0.1520**	**0.1545**	**0.1499**

few-shot model achieves a significant improvement over its zero-shot counterpart. In summary, the results demonstrate the effectiveness of reasoning methods, such as few-shot learning, which are incorporated into our approach and highlight the essential role of landscape knowledge for language models to perform domain-oriented reasoning.

5.2 Experiment 2: Illustration module performance

5.2.1 G1: Ensuring graphical coherence of the results generated during iterative design.

Quantitative evaluation. We first conduct a quantitative experiment to assess the efficacy of incorporating layout as a condition within rendering generation to sustain graphical coherence by comparing the similarity of images generated in multiple iterations.

Baseline model. For comparison, we create a baseline model by removing the layout-conditional module but keeping other model configurations including the same LoRA model. This ensures consistency in the knowledge level between the models and minimizes the risk of misinterpretations that could influence the experiment.

Data preparation. We develop the test dataset from two sources: randomly generated and expert generated, with considerations of two variables: plant type and plant number. First, we randomly generate 4800 landscape renderings including 5 compositions of landscape scenes that cover 20 different plant combinations: Tree A (C1), Tree A+Tree A (C2), Tree A+Tree B (C3), Tree A+Shrub A (C4), Shrub A (C5). Second, we recruit 7 designers with over three years of experience in landscape design. They are instructed to experience the interactive functionality of our system and use it to generate landscape renderings, fully simulating actual adjustments according to their professional experience. The generation process mentioned above employs the same prompt with the same seeds group. Each combination utilizes the same prompt and object layout while generated images differ according to random seeds.

Metric. We employ SSIM [19] to evaluate the images' graphical coherence during iterative generation. SSIM is a widely used metric for measuring the similarity between two images. It takes into account the structural information of the images instead of only the pixel values, which is more in line with human perception and suitable for the assessment of layout-guided generation models. The similarity of the generated images within the groups can be quantified by comparing the mean value of the SSIM. A greater SSIM score signifies a heightened coherence among the images generated under identical prompts, thereby reflecting the robustness of the model in iterative design.

Results. As shown in Table 2, our model consistently outperforms the baseline in all categories. The result signifies the effectiveness of introducing layout control in landscape rendering generation, particularly in enhancing image structural coherence throughout the iteration. Notable, the improvements are more significant in C2 & C3. This may be because plant types in C2 and C3 are both two trees, typically occupy a larger area than other groups in the image, making the consistency provided by the layout guidance more prominent across test sets.

Subjective evaluation. In landscape design, subjective perception of visual similarity can be affected by factors such as spatial ambiance, color and space harmony, which is not able to be evaluated using quantitative metrics. As a complement from subjective perspective, we further conduct a subjective user study to evaluate the graphical coherence considering human perception.
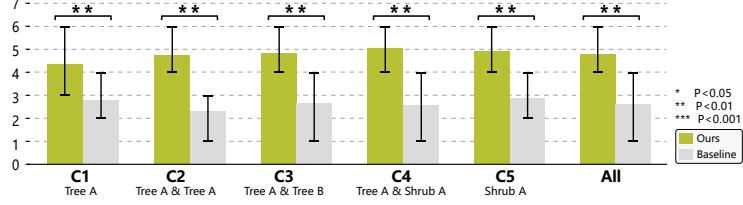
Fig. 9. Results of the user study comparing the graphical coherence of the generated images within the groups.

Participants. 45 participants are recruited by posting recruitment messages on online social media. Among them, 20 participants have design experience and the remaining 25 do not have. The average age of the participants is 24.9 (2.22% in [0,18], 53.33% in (18,25], 35.56% in (26,30], 8.89% in (31,40]).

Procedure. We randomly select 55 groups of generated images using each model with the same seed from the test set mentioned above. Participants are tasked to compare the graphical coherence of image groups, and give ratings form an 8-scale questionnaire. In the instructions given prior to the experiment, they are informed that 1) the questionnaire options represented perceptions ranging from "very dissimilar (0 scale)" to "very similar (7 scale)"; 2) the focus is on visual similarity, regardless of semantics; and 3) each question needs to be answered within 15 seconds.

Results. No outlier is identified from the ratings. We run the Friedman test on ratings for our approach and the baseline for each group. As demonstrated in Fig. 9 (*ALL*), our method outperforms the baseline model in terms of image coherence from a subjective perspective across all test groups. This is particularly evident in generating multiple objects, such as in groups *C2, C3* and *C4*, where a comprehensive understanding of the relationships between these objects becomes crucial for analysis. For single object generation, our method presents a narrower performance gap compared to the baseline model, possibly because the baseline model possesses the ability to generate a single object with high robustness. However, even in single object generation tasks, our method exhibits a significantly higher coherence compared to the baseline model. As the complexity of the generated images increases, the performance gap between our method and the baseline model widens. This indicates that our method is better suited for design tasks that require a higher level of coherence, especially when dealing with complex object relationships and compositions.

5.2.2 *G2: Ensuring the model's ability to comprehend landscape design descriptions.*

We conduct a user study to evaluate the models' comprehension of designers' intentions, which is facilitated by the layout-based conditional input and LoRA module in integrating plant-related knowledge into the generation process.

Baseline model. To eliminate the impact of the layout condition on the generated results, we select our layout-guided model without the LoRA module as the baseline model.

Data preparation. For test set preparation, we randomly generate 50 images from each model covering 20 plant combinations the same as above. The generated images are paired with corresponding generation seeds to create a 5-point Likert questionnaire. 10 landscape designers with more than three years of design experience are recruited to evaluate the alignment between design intent and images.

Results. As depicted in Fig. 10, our method exhibits a higher text-image alignment for plant type factors across all combinations, demonstrating that the LoRA module successfully integrates plant knowledge into the model and enhances its ability to accurately generate plant types. In terms of plant type, groups *C2, C3, C4* by our model all exhibit better performance than the baseline, indicating that our model is better at generating the appropriate plants in multi-object scenes. As shown in Fig. 11 (*02*), the baseline model fails to generate the correct plants. Specifically, *C5* has
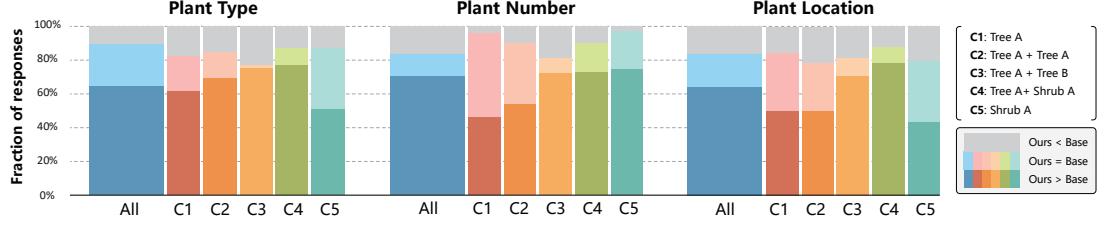
Fig. 10. Result of the user study comparing the model's comprehension of landscape design descriptions.
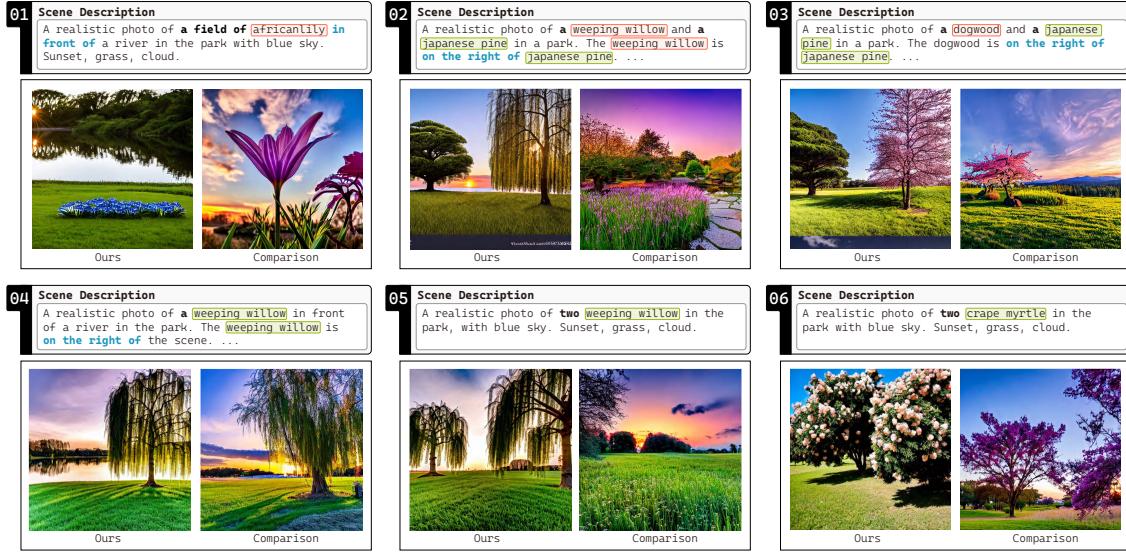

Fig. 11. Cases with a significant difference between the evaluation of our model and the baseline model.

the smallest gap between our model and the baseline model. This could be due to the fact that the shrubs in the baseline model overlap with the shrub species selected in our LoRA, resulting in both models performing well in this scenario.

Unexpectedly, our method shows a better comprehension of plant numbers and plant locations, demonstrating that the layout-based conditional input can efficiently enhance the model's comprehension of plant numbers and locations. In contrast, in complex scenes (C3, C4), our model is able to better control the correct number and location of plants. It is noteworthy that after adding plant species (C2/C3), the baseline model's ability to understand the number of plants decreased significantly, possibly due to the fact that pre-trained semantic information does not support such complex representations. Fig. 11 (*04,05*) reveals that the baseline model can comprehend *"a weep willow"* but fails to generate the correct scene for *"two weeping willow"*. In terms of plant location, our model still demonstrates superiority especially in complex scenes (*C3, C4*). As shown in Fig. 11 (*03*), the generated result omits the *"Japanese pine"* and locates the *"dogwood"* on the wrong side. These underline the need to include layout control in models to increase graphical coherence and produce visually captivating images.

17

6 USER STUDY OF *PLANTOGRAPHY*

To further evaluate the effectiveness of *PlantoGraphy*, we conducted a within-subjects study with 6 expert designers to compare the AI-supported landscape design process facilitated by *PlantoGraphy* with the conventional design process using widely utilized industry software.

6.1 Participants

We recruited 6 experts (*female*: 3 and *male*: 3, *Mean age* = 25.84) from social networks to participate in our study. To ensure a valid comparison, we selected participants who had experience using AI tools meanwhile had more than three years of experience with conventional design software. The demographics of all the participants are shown in Table 3. The participants were invited to conduct the experiments offline. Given the complexity of the experiment tasks, we covered the travel expenses for all participants and compensated them at a rate of about $13.5/hour.

Table 3. Demographics of the participants in system evaluation.

UID	Gender	Landscape design experience	Frequency on industry software	Frequency on AI tools
E1	Female	5 years	Very Frequently	Occasionally
E2	Female	7 years	Frequently	Rarely
E3	Male	7 years	Occasionally	Frequently
E4	Male	6 years	Very Frequently	Occasionally
E5	Male	5 years	Very Frequently	Occasionally
E6	Female	3 years	Frequently	Occasionally

6.2 Experiment Setup

For the user study, we devised a landscape design task simulating a real-world design creative process. The task was designed to assess participants' ability to utilize conventional designing tools *vs. PlantoGraphy* in creating landscape designs. The task involved three scenario descriptions as initial design requirements, as follows:

- Scenario 1: Wooded Path: A trail with tall trees on both sides.
- Scenario 2: Lake and Cherry Trees: The path bordered by a lake on one side and a row of cherry trees on the other, with some flowering plants around the lake.
- Scenario 3: Houses and Fields of Flowers: Small houses in the distance, with large fields of flowers nearby.

The participants were given the following task.

"You receive vague scenario descriptions from a client.
Prepare a concrete landscape rendering for the next debriefing session, to help confirm the project requirements."

6.3 Experiment Procedures

Prior to the study, the participants signed a consent form, agreeing to join the experiment, and allowing us to collect basic demographic information and record their behavioral data, including the frequency of operations on each function and the time taken to complete each task. Then, participants were instructed to go through the following steps:

(1) Participants were briefly introduced to the background of the project and received an introductory about the usage of *PlantoGraphy* interface for about 5 minutes. Then they were allowed to freely explore *PlantoGraphy* for about 10 minutes.

(2) Participants were asked to complete the design task using conventional landscape designing tools *vs.* using *PlantoGraphy*. For conventional design process, participants were provided with the flexibility to choose commonly
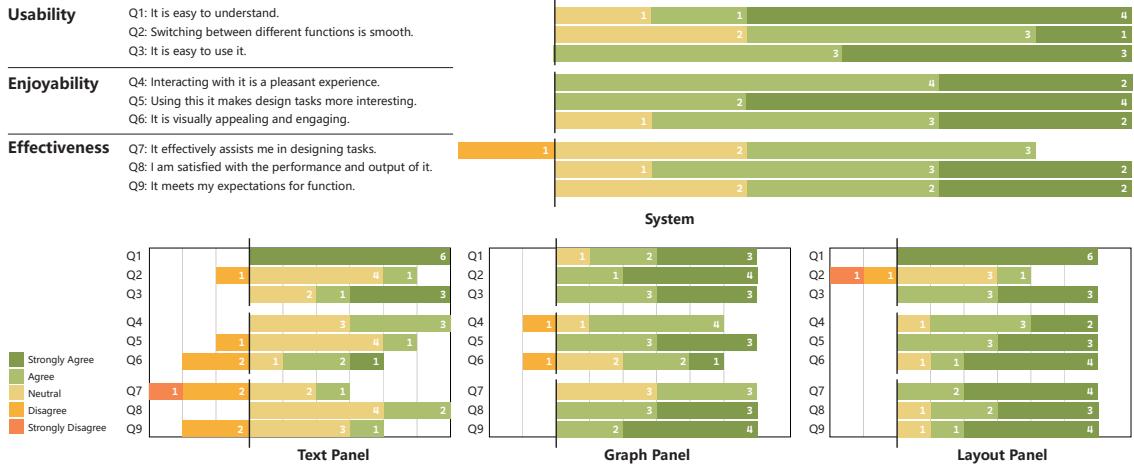
Fig. 12. Participant ratings on the *PlantoGraphy* system and the interactive panels in terms of *usability* (Q1-Q3), *enjoyability* (Q4-Q6), and *effectiveness* (Q7-Q9).
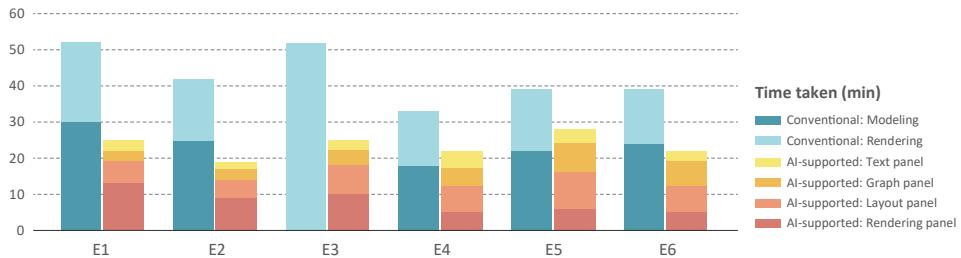

Fig. 13. Time taken on design tasks using conventional and AI-supported landscape design process by each participant.

used software that facilitates their design goals including modeling software (*e.g.*, Rhino, Sketchup, C4D) and rendering tools (*e.g.*, V-Ray, Lumion, Unreal Engine, Adobe Suite).

(3) Participants were asked to fill a post-study questionnarie with 5-point Likert scale questions focusing on *usability*, *enjoyability* and *effectiveness* ratings of each interactive panel and the whole *PlantoGraphy* system. Additionally, participants were asked to answer four open-ended questions:

- Does the interactive system improve the efficiency of collaborating with generative AI? If so, how?
- What is the difference between GAI-supported design process and traditional design process?
- What possible improvements can be made to the *PlantoGraphy* system?

To resolve bias by task familiarity, we randomized the order in which each participant using conventional tools or *PlantoGraphy*. We did not impose any time constraints to encourage participants to focus on the iterative design and prioritize the quality of their results. The average time for completing the study was 103.5 minutes for each participant.

6.4 Results

Below we first report the quantitative ratings and qualitative feedback from the participants on *system design*, followed by a *lessons learned* on AI-assisted design process.

6.4.1 System Design. User ratings regarding design of the overall system and interactive panels are depicted in Fig. 12.

- **Overall System.** Participants generally expressed satisfaction with the overall system's *usability* (Q1-Q3), *enjoyability* (Q4-Q6), and *effectiveness* (Q7-Q9). Specificly, the majority of users rated the system with the highest score in easy to understand (Q1) and to use (Q3). *E4 & E6* emphasized the importance of an intuitive interface for users engaging in design process. Most users enjoy using the *PlantoGraphy* system during creative process, with the highest score on making the design tasks more interesting (*Q5*). The participants unanimously lauded the system's interactive features, noting that "graphical editing panels align more effectively with typical design logic compared to conventional design process that requires several different softwares" (*E3*). Participant' response on the system effectiveness are relatively moderate. Notably, we can see a concern on assistance in designing tasks while some users gave a low score (*Q7*). Participant (*E6*) suggested that fine-tuning operations could be more intuitive if the layout were to be overlaid on the rendering panel. Additionally, *E1* observed that generating certain plant pairings, like weepwillow and banyan, can result in unusual edge formations in the plants. More discussions are provided in the lessons learned below.

- **Text Panel.** Participants showed a low level of interest on the Text Panel since the tasks were finalized and did not require any modifications. Interestingly, *E3* raised a high point in enjoyable for text panel: *"I like the weather and time slider that allow me to change the environment effect of generated image intuitively."* For improvement, participant suggested adding sketches as a more natural alternative for input, favoring their intuitive nature over traditional text-based methods (*E6*).

- **Graph Panel.** The ratings on the Graph Panel are mostly positive. *E2* found the graph representations and interactions especially beneficial, highlighting their potential to deepen understanding of complex structures. *"Plant combinations are influenced by various factors, including local climatic conditions, applicable plants, and the relationships between plant populations. ... Graph can abstract these relationships and it is intuitive (E2)". E3 & E6* suggested enhancements in the graph panel for smoother node addition and automated optimal node placement recommendations.

- **Layout Panel.** Participants perceived the Layout Panel to be the most enjoyable and efficient among the three panels. The layout panel was predominantly utilized, overshadowing the frequency of the graph and text panels, aligning with the subjective scoring provided by users. *E3* explained that *"designers prefer to use graphic-oriented approaches conveying design concepts with stakeholders." E5* also highlighted that "*using a graphical editing function for minor modification is more effective for iterative design*". However, *E4* gave the lowest score on switching between different functions (Q2), as the effect of layout modification was only visible on the output but not reflected on the other two panels.

Time Analysis. Fig. 13 illustrates the time spent by each participant on the tasks using conventional software and the *PlantoGraphy* system. Throughout the study, the traditional process demanded more time than AI-assisted ones. The reduction in time is quite significant (mean: 20.17 minutes, max: 27 minutes, min: 11 minutes), showcasing the efficiency of *PlantoGraphy* in streamlining design processes. Notably, the time reductions are more significant by E1~E3 than those by E4~E6. The follow-up interviews revealed the reasons, as E1~E3 took longer times for conceptualization that solely relies on designers' knowledge and experience using conventional tools, whilst GAI can quickly give some prototype designs and help designers brainstorm. Interestingly, E3 exclusively utilized Photoshop to merge online-retrieved plant images and create renderings. However, this approach proved time-consuming, as substantial effort was required to manually adjust plant sizes and positions to accurately portray human-perspective views.

We also observed distinct patterns in the temporal distribution of user engagement with the interactive panels in *PlantoGraphy* during creative tasks (Fig. 13). Here, time on the Rendering Panel pertains to the waiting durations for the model to generate renderings. Undoubtedly, participants devoted the least amount of time to the Text Panel and the most time to the Layout Panel, underscoring their inclination for iterative designs.

6.4.2 Lessons Learned. Participants highly praised the system's ability to support them in iterative design. *PlantoGraphy* provides opportunities for applying designers' expertise and collaborating with GAI, facilitating the development of distinctive and captivating designs. However, the improvement is constrained by *PlantoGraphy*'s limited flexibility in accommodating more diverse and complex design requirements. Participants attempted to incorporate additional plants that harmonized well with those in the given scenarios but encountered challenges, particularly when the number of plants exceeds six. This is likely attributable to the restricted quantity of plants in the fine-tuning dataset, and the LLM's reasoning capability when dealing with intricate graphs.

Upon conducting a detailed analysis of the time reduction disparity between E1~E3 and E4~E6, we discovered that E1~E3 primarily focused on the initial phase of the design process, whereas E4~E6 shifted their attention to the more advanced and in-depth design phase. E4~E6 noted that, "*for the more advanced stage, the content and quality of the images generated better matched my graphic sketches,*" necessitating additional iterative adjustments when using *PlantoGraphy*. The elongation of time in the traditional process is less pronounced, as the intermediate '3D model' outcome empowers designers with enhanced control for viewing and adjusting the design from various angles.

This observation reveals a compelling aspect of AI integration in design processes: as AI becomes more deeply embedded in the design workflow, its efficiency gains for designers seem to diminish. This phenomenon may stem from the escalating complexity of decisions and creative inputs required at advanced stages, a realm where AI-assisted design tools may not yet exhibit the same efficacy as in initial stage tasks. We anticipate that the progress of AI will enhance the effectiveness of AI-assisted design tools for more advanced tasks. Importantly, this also implies a potential reevaluation of the role of AI: from a tool that expedites tasks to one that evolves to enhance, rather than overshadow, human creativity and expertise.

7 DISCUSSION

7.1 Findings

7.1.1 Domain knowledge embedding. In our workflow, the end-to-end process of text-image generation is broken down into multiple steps, allowing domain knowledge to be introduced into the model development process in various forms: building prompt templates, translating experience into rules, etc. For example, the *concretization* incorporates designers' experience and knowledge into the domain-oriented LLM through prompt templates. An advantage is that we only need to construct a small number of datasets with experts and embedded the expertise into the model by means of few-shot learning, which is proved to be effective in Experiment 1 (Sect. 5.1). In this way, we utilize the powerful language processing and reasoning capabilities of LLMs to facilitate AI-assisted design. It requires less specifications of domain knowledge meanwhile retains high compatibility, which is friendly to non-AI practitioners.

7.1.2 Interaction design for generative AI. Our system introduces scene graph and layout as two intermediate graphical representations in the text-image generation process. These representations transform text-based scene descriptions into intuitive, easy-to-modify graphs from abstract and figurative perspectives. In post-study interviews, users highlighted the benefit of multimodal inputs (*e.g.*, text and graphical control) enabled by the GUI and widget interface, allowing for more

Exploration of a user   Renderings generated by a user
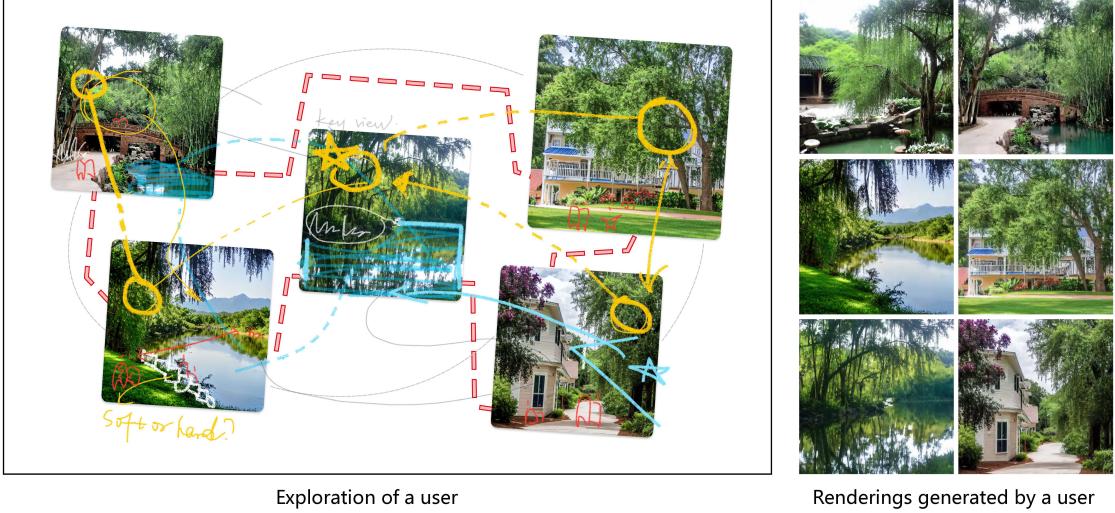
Fig. 14. Narrative map generated by a participant with the help of *PlantoGraphy*, illustrating his imagination of the human perspective key plots of the design site.

effortless and intuitive expression of design requirements. This facilitates the accurate translation and communication of the designer's needs to the generative AI models. Positive user feedback on interaction modes suggests that designing functions tailored to the target population's thought processes and operational logic reduces the learning curve and facilitates smoother information transfer. For example, designers in our study expressed familiarity with abstract graphs, aiding them in visualizing the system's potential for diverse design needs. These findings align with the guidelines for crafting human-centered generative AI systems as outlined in [40], mirroring findings from recent HCI-centric GAI studies for various domains (*e.g.*, [43, 50]).

7.1.3   Control from designers to models. Control from humans to models is regarded as a key dimension for human-centered generative AI systems [40]. In our approach, the use of layout as a conditional input enhances the graphical coherence of the generation model across iterations. Users found this improvement to be highly practical for real-world design scenarios, indicating that the AI tool effectively integrates into the iterative design process and co-creates with designers. Furthermore, controllable AI generation tools can empower designers to fully leverage their expertise and experience to stimulate creativity. Throughout the design process, designers can freely explore and experiment with various creative solutions, with AI serving as an assistive tool to offer additional possibilities and inspiration. In the future, AI generation tools incorporating multiple mediums of control will become increasingly crucial. Such tools will enable designers to achieve high-quality design outcomes more rapidly and enhance their work efficiency, all while capitalizing on the boundless creative potential offered by AI technology. However, it's essential to view AI tools as complementary, not substitutes, for professional landscape designers, as human expertise remains indispensable for ethical considerations, nuanced understanding, and client prioritization in landscape design.

7.1.4   AI-aided creativity. Many studies have demonstrated the capability of GAI systems in fostering creativity, by controlling parameters like random seed or penalty to create randomness in the outputs [30, 35]. *PlantoGraphy* also allow users to control the inputs to create diverse landscape renderings. Moreover, our system further introduces creativity by introducing multiple steps when utilizing GAI for landscape renderings. For instance, creativity enhancement

is evident during *brainstorming*, where *PlantoGraphy* allows designers to input basic environmental requirements and incrementally introduce plant combinations without a specific scene design. In addition, the ability to efficiently generating landscape renderings enables users to string these renderings together in an abstract streamline and form a narrative map, as shown in Fig. 14. In this novel AI-assisted design approach, the rendering, previously only present in the final result presentation, now serves as a substitute for the scene sketch and is incorporated into the initial concept development. We were pleased to discover that systems featuring intuitive and well-tailored interactive capabilities exhibited greater user-friendliness and were more readily embraced by designers, who could freely explore novel application scenarios. This realization lead us to consider that redefining the inputs and outputs of the traditional design process could serve as a starting point for envisioning a new design paradigm for creative design, particularly in light of the efficient generative capabilities offered by AI.

7.1.5 Method generalizability. While this study centers on landscape design, the proposed pipeline and fine-tuning strategies are readily adaptable to diverse creative designs. Breaking down an end-to-end process into a multi-step pipeline and integrating it with LLMs is applicable to other GAI-assisted designs with "iterating prompts and outputs" [43]. Specifically, a multi-step pipeline can enhance element relation arrangement and layout adjustment in the generation process, a characteristic often seen in architectural, fashion, and UI design. In these fields, textual descriptions alone may inadequately convey the intricacies of design intentions [26]. Our method, incorporating graph and layout as a structured forward process and interactive approach, enables a more nuanced interpretation of design intentions, simplifying the challenge of constructing a large-scale fine-tuning dataset. This encapsulation of domain knowledge into the pipeline proves essential for tasks where preparing a fine-tuning dataset is challenging.

7.2 Limitations

Generation performance. While in most cases, our system is capable of generating the correct type of plant in the intended location. However, there are instances where certain types of plants can influence each other. For example, the generated plant may be of the same type even if the user input specifies a different type. This issue may stem from an inherent limitation of GLIGEN. In our experiment, we observed that when certain objects are generated together, one object is converted into another within the GLIGEN base model. To address this situation, we attempt to mitigate the effect by using latent composition and freezing the generated single object latent. While this approach helps maintain the correct type of generated plant, it does result in a decrease in the coherence between the object and the background. In future work, we aim to investigate how and when different concepts cause this conversion to avoid such conditions and improve our system's usability.

Bias caused by model fine-tuning. Pre-trained models, while equipped with a broad knowledge base, can inadvertently carry biases if their training datasets lack diversity or exhibit certain preferences. We observed that while some popular styles like Japanese Zen Gardens and English Cottage Gardens are rendered with high fidelity, others are limited to a narrow selection of plant types or even incorrect species. Fine-tuning with LoRA is designed to customize the model's output more precisely for landscape design tasks. However, if the fine-tuning dataset is not sufficiently diverse, the model may become overly tailored to the particular styles, plants, and materials within that dataset, at the expense of other viable and innovative design options. This situation presents a critical trade-off: the specificity and accuracy of generated designs versus the diversity and versatility required by designers. To balance this trade-off, the weight of LoRA model, fine-tuning dataset collection and other methods like regularization should be carefully considered in accordance with the designers' objectives.

Another limitation is referred to as the concept pollution of LoRA. When a LoRA model is merged with the base model, certain concepts in the base model can be altered or contaminated by the LoRA model. For instance, if we load a LoRA model focusing on "banya" trees into the base model, all the trees in the base model may become "banya" trees. To alleviate this issue, we employ regularization techniques. However, it is worth noting that when encountering strong features, such as colorful flowers, it is possible for the flowers in the base model to be contaminated and appear as the same type of colorful flower. In future work, we aim to augment the regularization dataset to prevent such concept pollution and further enhance the system's performance.

System design. The current system's selection of plant species is limited due to the concept pollution problem. Resolving this issue could enable the system to support a wider range of species. Furthermore, plant selection in our system relies on the designer's experience for configuration, with the ecological harmony of configured plants judged solely by the designer, which raises the system's threshold. In the future, incorporating a graph recommendation model with expert knowledge input could enable automatic plant species recommendations and new graph generation on the graph panel, better aligning with landscape design practice.

AI-assisted system with multi-interactions. During the post-study interviews with designers, we learned that they are accustomed to using hand-drawn sketches and expressed their desire for a function allowing requirement transformation in the form of drawings directly on the drawing surface. This insight highlighted the importance of considering users' habits when selecting different interaction methods with the consideration of diverse user input (text, voice, drawing, graphical control, etc.). It can maximize the smoothness of user command transformation within the system.

8 CONCLUSION AND FUTURE WORK

This paper has explored the potential of integrating large pre-trained generative models within various design phases to facilitate a human-AI iterative design process. Addressing the limitations of existing end-to-end rendering generation methods, *PlantoGraphy* was developed through a formative study and support scene graph and layout as visual interaction components to ensure more compatibility with common design processes. The system incorporates a two-stage pipeline, consisting of a concretization module for translating conceptual ideas into concrete scene layouts, and an illustration module for transforming scene layouts into realistic landscape renderings. Performance evaluations have attested to *PlantoGraphy*'s effectiveness in landscape rendering generation. A within-subjects study comparing conventional design process with the AI-assisted design process has demonstrated the effectiveness of *PlantoGraphy* and also uncovered potential areas for improvements.

Future work will focus on enhancing the generation controllability of AI-assisted landscape designing tools by addressing current limitations, such as concept pollution and unexpected object type conversion. To augment user experience and efficiency, our system will also be equipped with supplementary features, such as plant pairing. Building upon this research foundation, we aim to delve deeper into novel design paradigms that foster collaboration between artificial intelligence and human designers. Additionally, we intend to investigate effective interaction patterns and pertinent information circuits tailored for the enhancement of future AI-supported design processes.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Oron Ashual and Lior Wolf. 2019. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4561–4569.

[3] Elizabeth Boults and Chip Sullivan. 2010. *Illustrated history of landscape design.* John Wiley & Sons.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[5] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.

[6] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. 2021. A Comprehensive Survey of Scene Graphs: Generation and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2021), 1–26.

[7] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. 2020. Sequential attention GAN for interactive image editing. In *Proceedings of the 28th ACM international conference on multimedia*. 4383–4391.

[8] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[9] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[10] Richard Lee Davis, Thiemo Wambsganss, Wei Jiang, Kevin Gonyop Kim, Tanja Käser, and Pierre Dillenbourg. 2023. Fashioning the Future: Unlocking the Creative Potential of Deep Generative Models for Design Space Exploration. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA, Article 136, 9 pages.

[11] Ellen Yi-Luen Do. 2005. Design sketches and sketch design tools. *Knowledge-Based Systems* 18, 8 (2005), 383–405.

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

[13] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. 2023. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986* (2023).

[14] Seamus W Filor. 1994. The nature of landscape design and design process. *Landscape and Urban Planning* 30, 3 (1994), 121–129.

[15] David Foster. 2023. *Generative Deep Learning, 2nd Edition*. O'Reilly Media, Inc.

[16] Gabriela Goldschmidt. 2014. *Linkography: unfolding the design process*. Mit Press.

[17] Daniel M Herbert. 1993. *Architectural study drawings*. John Wiley & Sons.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[19] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*. IEEE, 2366–2369.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[21] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[22] Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1219–1228.

[23] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3668–3678.

[24] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[26] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 919–933.

[27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *URL https://arxiv. org/abs/2205.11916* (2022).

[28] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

[29] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655* (2023).

[30] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376739

[31] Merriam-Webster. 2023. Definition of Generative AI. https://www.merriam-webster.com/dictionary/generative%20AI Accessed: 11 December 2023.

[32] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023).

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[34] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[35] Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy). 15–28.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[38] Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551* (2022).

[39] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057* (2022).

[40] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-Centric Survey and Taxonomy of Human-Generative-AI Interactions. *arXiv preprint arXiv:2310.07127* (2023).

[41] Yang Shi, Tian Gao, Xiaohan Jiao, and Nan Cao. 2023. Understanding Design Collaboration Between Designers and Artificial Intelligence: A Systematic Literature Review. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[43] Veera Vimpari, Annakaisa Kultima, Perttu Hämäläinen, and Christian Guckelsberger. 2023. "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-To-Image-Generation AI by Game Industry Professionals. *arXiv preprint arXiv:2302.12601* (2023).

[44] Bo Wang, Tao Wu, Minfeng Zhu, and Peng Du. 2022. Interactive image synthesis with panoptic layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7783–7792.

[45] Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916* (2023).

[46] Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. 2023. PointShopAR: Supporting Environmental Design Prototyping Using Point Cloud in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[47] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[49] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

[50] Di Wu, Zhiwang Yu, Nan Ma, Jianan Jiang, Yuetian Wang, Guixiang Zhou, Hanhui Deng, and Yi Li. 2023. StyleMe: Towards Intelligent Fashion Generation with Designer Style. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[51] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–10.

[52] Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. 2023. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[53] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. 2022. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138* (2022).

[54] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[55] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601* (2023).

[56] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

[57] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint arXiv:2210.03493* (2022).

[58] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

[59] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22490–22499.

[60] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. 2021. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Transactions on Cybernetics* 52, 8 (2021), 8574–8586.

A GLOSSARY

Term	Definition
Artificial Intelligence Generated Content (AIGC)	A new creation model incorporating artificial intelligence and human collaboration for content generation.
Bounding box	A rectangular box represents an object's spatial location and dimensional information.
Denoising Diffusion Implicit Model (DDIM) inversion	A technique is proposed for iteratively transforming data distributions, such as images, into random noise. It enables the model to reconstruct or generate specific instances of data, such as images, by retracing back its corresponding latent space in the diffusion process.
End to End	An idea that only focuses on inputs and outputs without considering intermediate processes.
Few shots	Another method to train models for specific tasks with a few samples.
Fine-tune	A technique to train a pre-trained model to optimize its performance on a specific task.
Gated Attention	It is a component of neural networks referred to as the attention mechanism that plays a crucial role in selectively focusing on targeted regions, thereby enhancing the model's ability to capture and extract key features.
Generative Artificial Intelligence (GAI)	Artificial intelligence can generate new content (e.g., text and images) through training on an extensive dataset [31].
Graphical User Interface (GUI)	An interface that provides visual information and allows interaction.
Large Language Model (LLM)	The language model can comprehend and produce human-like responses based on a large amount of textual data training.
Latent Diffusion Model	Based on the traditional diffusion model, this model applies the diffusion process to the latent encoding rather than the pixels, thus generating high-quality images more efficiently[28, 37].
Low-Rank Adaptation (LoRA) Diffusion Model	A method of fine-tuning large pre-trained models for text-to-image generation enables the training with fewer parameters while ensuring high performance, reducing computational cost, and allowing large models to excel in specific tasks [20].
Pre-trained text-to-image generation models	Generative artificial intelligence (GAI) models that have been previously trained can be further tuned to perform text-to-image tasks.
Prompt Engineering	One practice is to modify the prompts given to the large language models for better outputs that meet user needs.
Prompts	Instructions or requests given to generative artificial intelligence.
Stable Diffusion UNet	A symmetrical structure of the central component in the stable diffusion model plays a crucial role in effectively denoising images.
Zero shot	A training approach using a model directly to perform tasks without additional training.