

Deep Learning-Driven Generative Layout Design Model for Sensory Garden Environment Design

Jing Zhang 

College of Arts, Shandong Jianzhu University, Jinan, Shandong, China

*Corresponding author: Jing Zhang. Email: outlook_8A9201AAC2554F63@outlook.com

Abstract

The spatial layout design of sensory gardens involves the collaborative optimization of multiple design decisions. Traditional methods rely on manual experience and are inefficient in scheme exploration. This study proposes a diffusion generation model that integrates design grammar rules and conditional constraints to achieve end-to-end generation from land-use conditions to complete layout schemes. A multi-scale spatial feature encoder is designed to capture the spatial skeleton and element combination patterns of the garden through global- and local-scale hierarchical coding and cross-scale attention fusion. A conditional embedding module is constructed, in which hard constraints such as the land-use red line and setback distance are encoded as differentiable vectors, and soft constraints such as visual permeability and spatial enclosure are encoded as differentiable vectors. A graph convolutional network is used to explicitly model the adjacency, inclusion, and axis relationships among design elements, and the topological prior is injected into the denoising process. The U-Net backbone is improved, and a dual-channel attention mechanism and a progressive refinement strategy are introduced to enhance structural awareness and boundary accuracy. Experiments are carried out on a dataset containing 200 cases, and ten indicators such as hard constraint violation rate, Fréchet distance, and spatial enclosure are used for evaluation. The results show that the hard constraint violation rate of this model is 2.3%, which is 84% lower than that of the standard diffusion model. The Fréchet distance is 8.7, and the spatial enclosure is 59.3%. All ten indicators outperform the four baseline methods such as LayoutGAN and LayoutVAE. The ablation experiment confirms the independent contribution of each core component, and the cross-shape generalization and small-sample experiments verify the adaptability and data efficiency of the model, which provided an effective method support for generative landscape design.

Keywords

Deep learning; Diffusion model; Generative design; Sensory garden; Design of environment

Received: 30 Jun 2026; Revised: 26 Jul 2026; Accepted: 24 Aug 2026; Published: 01 Sep 2026
Copyright: ©2026 The Author(s). Published by [Digital Intelligence Press Limited](#). This is an
open access article under the [CC BY 4.0](#) license.

1. INTRODUCTION

As a composite outdoor spatial environment with plant landscape as the main body and integrating water features and hard landscape elements, the core design goal of a sensory garden is to create a landscape space with a rich spatial experience through the organic organization of multi-sensory stimulation elements such as vision, hearing, smell, and touch [1],[2]. The spatial layout design of this kind of garden involves the collaborative optimization of multiple design decisions, such as functional zoning, vegetation configuration, garden path organization, water feature form, and landscape composition. There are complex spatial relations and functional dependencies among the design elements, and the layout scheme directly determines the spatial quality and user experience of the

garden [3]. However, current sensory garden design practice mainly relies on the personal experience and manual deliberation of designers, and the process from concept sketch to final scheme requires repeated modification and iteration, which is not only inefficient in design but also makes it difficult to fully explore the diversified possibility space within the limited design cycle. Although traditional parametric design methods can assist in generating rule-based layout schemes to a certain extent, the diversity of their generation is limited by the completeness of preset rules, and they lack effective quantitative expression capability for the qualitative design principles and fuzzy spatial boundaries that exist in sensory gardens. The case-based reasoning (CBR) design method faces the dual dilemma of limited case base size and insufficient retrieval and matching accuracy, which makes it difficult to generate innovative layout schemes under unfamiliar design conditions [4],[5].

In recent years, deep learning technology has made breakthroughs in the field of generative design. Deep generative models represented by generative adversarial networks, variational autoencoders, and diffusion models have shown the ability to automatically learn spatial distribution rules from large-scale data and to generate novel design schemes in building layout generation, urban block form design, indoor space configuration optimization, and other tasks. Among them, the diffusion model, with its progressive denoising generation mechanism, shows significant advantages over traditional generation paradigms in terms of generation quality, training stability, and pattern coverage breadth, and has become a cutting-edge method in the field of structured layout generation [6],[7]. However, there are two key challenges when directly transferring the diffusion model to the sensory garden layout generation task. First, the layout of sensory gardens contains rich spatial topological relations, with adjacency relations defining spatial continuity, inclusion relations encoding spatial hierarchy, and axis relations governing spatial sequence order. Standard diffusion models only take pixel-level grids as processing objects, so they lack the ability to explicitly model these spatial relations [8]. As a result, the generated scheme is prone to structural problems such as element boundary interpenetration and axis relationship disruption. Second, there are a large number of constraints from land-use conditions, regulatory requirements, and design principles in actual engineering design, including mandatory constraints such as the land-use red line, setback distance, and fire access, as well as optimization constraints such as visual permeability, spatial enclosure, and area ratio. The standard diffusion model does not have a mechanism to inject these design constraints as generation conditions into the generation process. The generated schemes are often visually desirable but flawed in terms of engineering feasibility.

To solve the above problems, this study proposes a sensory garden layout generation model that integrates design grammar rules and conditional constraints, aiming to build an end-to-end generation framework from the input of design conditions to the output of a complete layout scheme [9],[10]. The core innovations of the model are reflected in three aspects. First, a multi-scale spatial feature encoder is designed to capture the overall land contour and spatial sequence skeleton of the garden through global-scale coding, perceive the element combination patterns inside each functional zone through local-scale coding, and achieve full interaction of cross-scale information through pyramid pooling and a spatial attention mechanism, so that the model has multi-level spatial understanding capability from the whole to the local. Second, a conditional embedding module that integrates design constraints is constructed to encode hard constraints such as the land-use boundary, setback requirements, and fire access, as well as soft constraints such as visual permeability, spatial enclosure, and functional zone area ratios, into differentiable condition vectors, which are injected into the denoising network via adaptive layer normalization. At the same time, a graph convolutional network is introduced to explicitly encode the contiguity, inclusion, and axis relationships among design elements, and the spatial

topological prior is embedded in the feature representation of each layer during the generation process. Third, the denoising network backbone of the diffusion model is improved, a dual-channel attention module is introduced into the standard U-Net architecture to enhance selective attention to key spatial regions and semantic channels, and a progressive refinement strategy from coarse-grained functional partition generation to fine-grained design element positioning is designed, so that the generation process simulates the design thinking process of human designers from overall layout to local refinement. Through the synergy of the above three improvements, the model can maintain the high-fidelity generation capability of the diffusion model while having endogenous compliance capability with design grammar rules and engineering constraints.

2. CODING OF ENVIRONMENTAL DESIGN ELEMENTS AND CONSTRUCTION OF DATASETS FOR SENSORY GARDENS

2.1 Deconstruction of design elements of the spatial layout of the sensory garden

The spatial organization of a sensory garden follows the basic principles of landscape design, and its layout structure can be divided into several functional zones with distinct characteristics according to their function and spatial attributes. In this study, the planar space of the sensory garden is divided into four types of core functional zones. The entrance area is responsible for the starting and transitional functions of the spatial sequence and is usually adjacent to the external road or the main flow of people. Its spatial form is mainly open or semi-open, and its area generally accounts for 8% to 15% of the total area of the garden. The rest area is a semi-private space for users to stop and sit for viewing. It is often equipped with seating, pavilion and other facilities for staying. The space enclosure degree is high, accounting for about 20% to 30% of the area. The viewing area is the most concentrated area of plant landscaping in the sensory garden, with the main purpose of displaying flower color, plant texture and seasonal phase change. The spatial form is mostly flake-like or ribbon expansion, and the area proportion is usually 35 to 50 percent. The traversal area takes the park road system as the main body, which undertakes the traffic function of connecting various functional zones and organizing tour lines [11],[12]. Its width varies according to the grade of the main and secondary park roads. The width of the main park road is generally 1.5m to 3m, and the width of the secondary park road is 0.8m to 1.5m.

Under the framework of the above functional zoning, the design elements of the sensory garden can be divided into three basic types according to the geometric dimensions: points, lines and planes. Each element has clear geometric characteristics, scale range and spatial function positioning. [Table 1](#) summarizes the classification system and geometric parameters of the seven types of core design elements.

Table 1. Classification system and geometric parameters of sensory garden design elements

Category of elements	Name of element	Dimension of geometry	Typical scale range
Element of point	Landscape Sketch	Dimension of zero	1-3 m in diameter
Element of point	Planting trees alone	Dimension of zero	The crown width is 3-8 m
Element of line	The Garden Road	One dimension	Width 0.8-3.0m
Element of line	Green hedge	One dimension	Width 0.5-1.5m
Element of surface	Lawn	Two dimensions	Variable area
Element of surface	Surface of water	Two dimensions	Variable area
Element of surface	Landscape of flowers	Two dimensions	Variable area

Point elements include landscape sketches and isolated trees. Landscape pieces such as sculpture, stone, water bowl, etc., have a point-like plane projection with a diameter of 1 to 3 meters, which acts as visual focus and spatial marker in space. Solitary trees are independent ornamental trees with a canopy range of 3 to 8 meters, and their location and scale control the spatial composition of the whole garden. Line elements include garden road and hedge. The park road system includes main roads and secondary tourist trails, with a width of 0.8 to 3.0 meters, which undertakes the function of traffic organization and tour line guidance. As a ribbon planting belt, the hedgerow has a width ranging from 0.5 to 1.5 meters, which serves as both spatial separation and visual guidance. Surface elements include lawn, water surface and flower landscape. Lawns serve as a large area of open ground cover, providing visual penetration areas. The water surface includes natural and regular water features, which regulate the microclimate and form a visual focus through evaporation and heat absorption. The flower habitat is composed of a variety of flowers and ornamental grasses, which form a patch planting community. It is the most abundant design element in the sensory garden to display color change and seasonal phase. The above seven types of elements are organized into a complete layout scheme through three types of spatial relations: contiguity, inclusion and axis.

2.2 Graph representation and vector coding method of design scheme

In order to transform the design scheme of the sensory garden into the numerical input that can be handled by the deep learning model, this study adopts the dual coding strategy combining graph structure and grid matrix to correspond to the topological relationship expression and spatial layout expression of the design scheme respectively.

Graph-based scene representation abstracts the sensory garden design scheme into an undirected graph with attributes [13]:

$$G = (V, E, A, R) \quad (1)$$

Where V is the set of design element nodes, E is the set of spatial relation edges, A is the attribute feature matrix of nodes, and R is the type coding matrix of edges. For any node $v_i \in V$, its attribute vector a_i contains the spatial coordinates, geometric scale and category information of the element, and the specific form will be given in the vector coding section.

For the edge $e_{ij} \in E$ connecting nodes v_i and v_j , its type $r_{ij} \in \{\text{adjacent, containing, axis}\}$ corresponds to the three spatial relations defined in the previous section. The advantage of graph structure coding is that it can explicitly retain the topological connection relationship between design elements and has the flexibility to adapt to any number of elements, which is not limited by fixed resolution raster.

Vectorized layout parameterizes each design element with a complete set of geometric attribute values. For any element node i , its parameter vector is expressed as:

$$p_i = (x_i, y_i, c_i, s_i, \theta_i, t_i) \quad (2)$$

Where x_i and y_i are the horizontal and vertical coordinates of the geometric center of the element in the garden plane coordinate system, the unit is meters, and the origin of the coordinate system is the southwest corner of the garden land.

c_i is the category label of the elements, which is expressed by unique thermal coding. The specific corresponding relations are as follows: landscape piece is denoted as category 1, isolated trees are denoted as category 2, garden road is denoted as category 3, hedgerows are denoted as category 4, lawns are denoted as category 5, water surface is denoted as category 6, flower landscape is denoted as category 7;

s_i is the scale parameter of the element. For the point element s_i , the equivalent diameter is taken, for the line element s_i , the width value is taken, and for the surface element s_i , the square value of the area is taken. θ_i is the principal direction angle of the element, which is expressed in radian system, and the range is $[-, \pi\pi]$. For isotropic point elements, the value is zero; t_i is the type of modification parameter of elements, which is used to distinguish different design subcategories in the same type of elements. For example, the garden road is further divided into the main garden road and the secondary garden road, and the water surface is further divided into the natural type and the regular type.

Through the above vector coding method, the complete layout information of each design scheme is converted into an ordered set with an indefinite number of nodes but fixed attribute dimensions, which is convenient for the subsequent access of graph neural network and converter model.

The rasterized layout matrix coding discretizes the continuous plan of the sensory garden into a fixed size raster grid to construct a multi-channel semantic segmentation map. Let the garden land size be $W \times H$ (in meters) and the rasterization resolution be $N \times N$, then the actual ground size corresponding to each pixel unit is:

$$\Delta = \frac{W}{N} = \frac{H}{N} \quad (3)$$

For each pixel unit, channel coding is carried out according to the design element category to which it belongs, and a $C \times N \times N$ tensor M is generated, where C is the total number of design element categories are. In this paper, $C = 7$ specific coding rules are as follows: for the k -th category channel, if the pixel position (m, n) belongs to the coverage range of design element of this category, then:

$$M_k(m, n) = 1 \quad (4)$$

$$M_k(m, n) = 0 \quad (5)$$

Different types of design elements do not overlap in space (except for the possibility that the hedge and the garden road may be parallel and adjacent rather than overlapping), so the value of each channel

at the same pixel location is mutually exclusive.

The advantage of raster coding is that it naturally ADAPTS to the input format of convolutional neural network and facilitates the extraction of spatial texture and neighborhood features, but its resolution is limited by the selection of N [14]. In this paper, when $N = 256$, the spatial resolution corresponding to a single pixel is $\frac{W}{256} \times \frac{H}{256}$, which is enough to express the contour details of the design elements. Graph representation and raster representation are converted to each other through spatial coordinate mapping. The former serves the topological constraint modeling of the generative model, while the latter serves the visual feature learning of the generative model. The two together constitute the dual perspective representation of the design scheme.

2.3 Data set construction and preprocessing

The size and quality of data set directly determine the generation ability and generalization performance of deep learning model. In this study, sensory garden design cases were collected from multiple channels to construct a dedicated dataset. [Table 2](#) summarizes the basic statistical information such as the original collection quantity, the effective retention quantity, the average land area and the average number of design elements for each data source type.

Table 2. Basic statistics of the Sensory Garden design dataset

Type of data source	Original number of cases	Effective reservation number	Average land area (m ²)	Average number of factors (units)
Physical gardens have been built	62	51	1850	18.4
Design the winning proposal for the competition	118	92	1240	22.7
Open database and design institute information	76	57	2080	15.9
Total/weighted average	256	200	1630	19.2

Data sources mainly include three types of channels. The completed physical garden cases obtained the layout information through field mapping and aerial images, covering the representative sensory garden projects at home and abroad. Sixty-two were originally collected, and 51 were effectively retained after quality screening. The land area of this kind of case is the largest, with an average of 1,850 square meters, reflecting the typical land scale of sensory garden in the actual project. The winning proposals of the design competition were collected from IFLA student competition, CHSLA student competition and other channels. 118 original proposals were collected, and 92 were effectively retained. The average land area of this kind of case was 1,240 square meters, which was the smallest among the three categories, but the average number of design elements reached 227, which was the highest among the three categories. It reflects the characteristics of the competition scheme to pursue higher design complexity and element richness within the limited land. The public database and the data of the design institute include the project database of well-known landscape design companies at home and abroad. Seventy-six were originally collected and 57 were effectively retained. The average land area is 2,080 square meters, the largest among the three categories, but the average number of elements is only 15.9, indicating that the actual project is more concise and restraint than the competition plan. A total of 256 original cases were collected from the three types of sources. After cleaning steps such as image registration, edge detection vector-ization, design element annotation and spatial scale

normalization, 200 valid samples were retained. The weighted average land area of all samples was 1,630 square meters, and the average number of elements was 19.2.

2.4 Formal definition of design constraints

The generative design of sensory gardens must meet the multiple constraints from land use conditions, normative requirements and design principles. In this study, design constraints are divided into two categories: hard constraints and soft constraints.

Hard constraints are mandatory conditions that must be met unconditionally by the generation plan, and if any of them is violated, it will be judged as invalid, including three types of land use red line constraints, boundary withdrawal requirements and fire access requirements. The red line constraint of land use stipulates that all design elements must fall completely within the red line boundary Ω of garden land use, that is, the occupied area A_i of any design element v_i must meet:

$$A_i \subseteq \Omega \quad (6)$$

The rasterization is expressed as:

$$M_k(m, n) \leq B(m, n), \forall k, \forall (m, n) \quad (7)$$

The withdrawal requirement specifies that the shortest Euclidean distance between the geometric center of point elements (x_i, y_i) and the land use boundary is not less than the minimum concession distance d_{setback} (the value is 1.5 to 3.0 m), namely:

$$\min_{q \in \partial\Omega} \|(x_i, y_i) - q\|_2 \geq d_{\text{setback}} \quad (8)$$

Any point on the contour of planar element must also satisfy this constraint. Fire access requirements require that the garden should retain a continuous passage with a net width of no less than 4 m, which does not cover the occupied area of any point elements or surface elements, and can be numerically verified by morphological expansion and connectivity detection on the grid diagram.

The soft constraint is the optimization condition that the design scheme should meet as much as possible but allow a certain degree of deviation. The violation of the soft constraint does not lead to the invalidation of the scheme but will reduce the comprehensive quality score, including the three items of visual field permeability, spatial enclosure and the ratio of functional zone area. V_{openness} is defined as the ratio of the average visible distance of the observation point in all directions to the characteristic scale of the garden:

$$V_{\text{openness}} = \frac{1}{|S_{\text{view}}|} \sum_{s \in S_{\text{view}}} \frac{\sum_{\phi \in \Phi} d_{\text{visible}}(s, \phi)}{L_{\text{char}} |\Phi|} \quad (9)$$

Where S_{view} is the set of uniformly sampled observation points in the grid space, Φ is the discrete set of eight principal azimuth angles, $d_{\text{visible}}(s, \phi)$ is the visible distance from the observation point s to the ray emitted along the azimuth angle ϕ until it meets the occlusion element or the land boundary, the garden equivalent diameter is:

$$L_{\text{char}} = \sqrt{WH} \quad (10)$$

The spatial enclosure degree $C_{\text{enclosure}}$ is defined as the proportion of the area perimeter of the surface element occupied by the boundaries of other design elements:

$$C_{enclosure} = \frac{1}{N_{face}} \sum_{i=1}^{N_{face}} \frac{L_{adjacent}(A_i)}{L_{perimeter}(A_i)} \quad (11)$$

Where N_{face} is the total number of surface elements, $L_{perimeter}(A_i)$ is the total perimeter of the i -th surface element region, and $L_{adjacent}(A_i)$ is the length of the boundary shared with adjacent elements in the region perimeter. The area ratio deviation of functional partitions D_{ratio} is defined as the sum of the squares of the differences between the actual area ratio and the recommended area ratio of the four core partitions:

$$D_{ratio} = \sum_{r \in R} \left(\frac{A_r}{A_{total}} - p_r \right)^2 \quad (12)$$

Among them:

$$R = \{\text{The Entrance, Take a break, Watch, Go through}\} \quad (13)$$

The corresponding recommendation ratios p_r are 0.12, 0.25, 0.42, and 0.21, respectively. Hard and soft constraints are embedded in model training in the form of differentiable penalty terms. Hard constraint penalty item:

$$L_{hard} = \lambda_{hard} \cdot v_{hard} \quad (14)$$

Where v_{hard} is the sign of hard constraint violation (1 for violation of any condition, 0 otherwise), λ_{hard} is 10.0. Soft constraint loss is defined as the weighted sum of three terms:

$$L_{soft} = w_V \max(0, V_{target} - V_{openness}) + w_C \max(0, C_{target} - C_{enclosure}) + w_D D_{ratio} \quad (15)$$

Among them:

$$V_{target} = 0.45, C_{target} = 0.55 \quad (16)$$

Is the expected threshold, and the weight coefficient is:

$$w_V = 0.5, w_C = 0.3, w_D = 0.2 \quad (17)$$

The values are 0.5, 0.3 and 0.2, respectively. In section 3, the above constraint penalty term will cooperate with the denoising loss of diffusion model to participate in gradient transmission and guide the generative model to gradually learn the implicit compliance ability to design constraints.

3. CONDITIONAL GENERATION MODEL ARCHITECTURE INTEGRATING DESIGN SYNTAX RULES

3.1 Multi-scale spatial feature encoder

The layout structure of the sensory garden presents a significant scale hierarchy, from the spatial skeleton defined by the overall land outline to the organizational order of each functional zone, and then to the precise location and scale of individual design elements [15]. The spatial characteristics at different scales jointly determine the visual quality and functional rationality of the design scheme. To this end, this study designs a multi-scale spatial feature encoder to extract and fuse layout information

from global, local and cross-scale levels.

The global scale coding focuses on the spatial structure characteristics of the garden at the overall level. This coding branch takes the garden's land boundary mask and spatial sequence marker as input, and the land boundary mask is a binary matrix [16]:

$$B \in \{0, 1\}^{N \times N} \quad (18)$$

Where N is the grid resolution, the pixel with the value of 1 means that it is within the red line of land use, and the pixel with the value of 0 means that it is outside the red line of land use. The spatial sequence is marked with a vector of variable length:

$$S_{\text{seq}} = [s_1, s_2, \dots, s_T] \quad (19)$$

Where T is the number of key nodes in the spatial sequence, and the value of each element s_k is the coordinate index of the k -th important spatial node in the garden, which usually corresponds to the entrance center, the turning point of the main axis, the main landscape focus and the end point of the spatial sequence. The global scale encoder uses a lightweight convolutional neural network to downsample the land boundary mask into a global feature map [17],[18]:

$$F_{\text{global}} \in \mathbb{R}^{C_g \times \frac{N}{8} \times \frac{N}{8}} \quad (20)$$

Where C_g is the number of global feature channels, set to 128. This branch aims to capture the overall shape characteristics of the garden, the spatial axial trend and the macro pattern of land use.

Local scale coding focuses on the pattern of element combination within each functional zone. The garden plane is divided according to the functional partition label, and the distribution slices of design elements inside the four partitions are obtained respectively. For each partition:

$$r \in \{\text{entrance, rest, view, walk through}\} \quad (21)$$

Take the corresponding local raster slice:

$$M_r \in \{0, 1\}^{C \times \frac{N}{2} \times \frac{N}{2}} \quad (22)$$

Where C is the number of design element categories and $N/2$ is the typical size of the partition (the size of each partition varies with the actual area and is unified to $N/2$ resolution through RoI Align). The local encoder uses the convolutional network with shared weights to independently code the combination of elements in each partition and obtain the local features:

$$F_{\text{local},r} \in \mathbb{R}^{C_l \times \frac{N}{4} \times \frac{N}{4}} \quad (23)$$

C_l is set to 64. The local features of the four partitions are spliced along the channel dimension and compressed into a unified local feature graph by the aggregate convolution layer:

$$F_{\text{local}} \in \mathbb{R}^{C_l \times \frac{N}{4} \times \frac{N}{4}} \quad (24)$$

The value of local scale coding is to enable the model to learn the internal design patterns of different functional zones, such as the combination law of flower landscape and point trees in the viewing area, and the configuration relationship between seating facilities and shade trees in the rest area.

Cross-scale feature fusion modules solve the problem of information interaction and scale alignment between global and local features. This study uses a fusion strategy combining spatial pyramid pooling with attention mechanism. Spatial pyramid pooling Firstly, the global feature map F_{global} is pooled at three different scales. The pooling window sizes are 2×2 , 4×4 and 8×8 respectively, and the context feature vectors at three scales are obtained, which are restored to a unified resolution of $N/8 \times N/8$ after upsampling. The multi-scale context feature map F_{pyramid} attention fusion module calculates the spatial attention weight of the global context to the local feature, so that the local feature can adaptively adjust its expression according to the global layout state. Specifically, the attention query tensor is generated by 1×1 convolution of global feature F_{global} :

$$Q_g \in \mathbb{R}^{\frac{N}{8} \times \frac{N}{8} \times C_{\text{attn}}} \quad (25)$$

The local feature F_{local} is bilinear upsampling and convolution with 1×1 to generate the key tensor $K_l \in \mathbb{R}^{\frac{N}{8} \times \frac{N}{8} \times C_{\text{attn}}}$ and the value tensor $V_l \in \mathbb{R}^{\frac{N}{8} \times \frac{N}{8} \times C_{\text{attn}}}$, where C_{attn} is the number of attention channels is set to 64. The calculation formula of spatial attention weight matrix $A \in \mathbb{R}^{(\frac{N}{8} \times \frac{N}{8}) \times (\frac{N}{8} \times \frac{N}{8})}$ is as follows:

$$A_{i,j} = \frac{\exp\left(\frac{Q_g^{(i)} \cdot K_l^{(j)}}{\sqrt{C_{\text{attn}}}}\right)}{\sum_k \exp\left(\frac{Q_g^{(i)} \cdot K_l^{(k)}}{\sqrt{C_{\text{attn}}}}\right)} \quad (26)$$

Where i and j represent the spatial position index in the global feature graph and the local feature graph respectively, $Q_g^{(i)}$ is the query vector of the global feature graph at position i , $K_l^{(j)}$ is the key vector of the local feature graph at position j , and $\sqrt{C_{\text{attn}}}$ is the scaling factor to maintain the gradient stability. Output features of attention fusion:

$$F_{\text{fused}} \in \mathbb{R}^{C_f \times \frac{N}{8} \times \frac{N}{8}} \quad (27)$$

It is calculated as follows:

$$F_{\text{fused}} = \text{Conv}([F_{\text{global}}, AV_l]) \quad (28)$$

Where $[\cdot, \cdot]$ represents the splicing operation on the channel dimension, Conv is 3×3 convolutional layer, C_f is the number of fused feature channels, set to 192. After cross-scale fusion, the output features of the encoder carry both the overall spatial skeleton information of the garden and the combination details of elements inside each partition, providing sufficient multi-scale semantic input for the subsequent diffusion denoising process.

3.2 Design of conditional embedding module

The design generation process of the sensory garden must meet a series of constraints defined by the design code and landscape principles. The task of the conditional embedding module is to transform these constraints into numerical guidance signals and inject denoising networks to control the generation direction. The module consists of three sub-modules: design constraint coding, adaptive condition injection and spatial relationship prior embedding.

The coding of design constraints converts the hard and soft constraints defined in Section 2.4 into a dense vector representation. The hard constraint is a binary condition, and the dimension N_{hard} of its coding vector $h \in \{0, 1\}^{N_{\text{hard}}}$ is taken as the maximum coding length of each item in the three hard constraints. The red line of land use is encoded as a binary matrix M_{boundary} of $N_{\text{bin}} \times N_{\text{bin}}$. If $N_{\text{bin}} = 16$, the land use boundary is downsampled to 16×16 resolution. The unbounded constraint is encoded as a scalar value:

$$d_{\min} = \min_i \min_{q \in \partial\Omega} \|p_i - q\|_2 \quad (29)$$

Represents the minimum distance between all elements and land boundaries; The fire channel constraint is coded as a binary variable $v_{\text{fire}} \in \{0, 1\}$ indicating whether the fire channel meets the net width requirement. The final coding vector of the hard constraint is the dense vector obtained by linear transformation after splicing of the above three codes:

$$c_{\text{hard}} \in \mathbb{R}^{D_{\text{hard}}} \quad (30)$$

Among them:

$$D_{\text{hard}} = 64 \quad (31)$$

The soft constraint is a continuous value condition, and its coding vector is:

$$c_{\text{soft}} \in \mathbb{R}^{D_{\text{soft}}} \quad (32)$$

It is directly composed of the current values of the three soft constraint indicators:

$$c_{\text{soft}} = \begin{bmatrix} V_{\text{openness}} \\ C_{\text{enclosure}} \\ D_{\text{ratio}} \end{bmatrix} \quad (33)$$

The definitions of V_{openness} , $C_{\text{enclosure}}$ and D_{ratio} are the same as those in section 2.4 of section 2.

$$D_{\text{soft}} = 3 \quad (34)$$

The hard and soft constraint codes are concatenated into the total condition vector:

$$c_{\text{cond}} \in \mathbb{R}^{D_{\text{cond}}} \quad (35)$$

Among them:

$$D_{\text{cond}} = D_{\text{hard}} + D_{\text{soft}} = 67 \quad (36)$$

The conditional fusion strategy uses adaptive layer normalization to inject the conditional vector into the denoising network of the diffusion model. Adaptive layer normalization is different from traditional batch normalization or layer normalization. Its normalization parameters (scaling factor γ and offset β) are dynamically generated by the conditional vector through learnable affine transformation, so that the network activation value can be adaptively adjusted according to the conditional input.

For the feature tensor $X^{(l)}$ of layer l in the denoising network, the calculation process of adaptive layer normalization is as follows:

$$\mu = \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} X_{h,w}^{(l)} \quad (37)$$

$$\sigma^2 = \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} (X_{h,w}^{(l)} - \mu)^2 \quad (38)$$

$$\gamma_l = W_\gamma^{(l)} c_{\text{cond}} + b_\gamma^{(l)} \quad (39)$$

$$\beta_l = W_\beta^{(l)} c_{\text{cond}} + b_\beta^{(l)} \quad (40)$$

$$\hat{X}^{(l)} = \gamma_l \frac{X^{(l)} - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta_l \quad (41)$$

Where H_l and W_l are the height and width of the feature map of layer l , μ and σ^2 are the mean and variance of the feature map of this layer, and $\varepsilon = 10^{-5}$ is a numerically stable small constant. γ_l for vector of condition with β_l c_{cond} via linear layer $W_\gamma^{(l)} \in \mathbb{R}^{C_l \times D_{\text{cond}}}$ and $W_\beta^{(l)} \in \mathbb{R}^{C_l \times D_{\text{cond}}}$ calculated by channel modulation parameters, C_l layer on the number of channels for the first l . Through adaptive layer normalization, conditional constraint signals are injected into each layer of the denoising network to continuously guide the generation process to evolve towards satisfying the design constraints during the whole forward propagation process.

Spatial relationship prior embedding explicitly encodes the topological relationship between design elements into the model feature space through graph convolution module [19],[20]. The module has the graph structure defined in section 2.2. As input, a multi-layer graph convolutional network is used to update the neighborhood aggregation of node attributes. The l -layer update rule of graph convolution is as follows:

$$a_i^{(l)} = \text{AGGREGATE}^{(l)} \left(\{h_j^{(l-1)} : j \in N(i)\} \right) \quad (42)$$

$$h_i^{(l)} = \text{COMBINE}^{(l)} \left(h_i^{(l-1)}, a_i^{(l)} \right) \quad (43)$$

Where $h_i^{(l-1)}$ is the feature vector of the i -th node in the previous layer, $N(i)$ is the first-order neighborhood set of node i in the graph, AGGREGATE is the neighborhood aggregation function, this paper adopts the weighted aggregation method of taking the feature mean of the neighborhood node plus the edge type coding, COMBINE is the combination function. The gated cyclic unit is used to fuse the node's own features with the neighborhood aggregation features. After the convolution update of $L_{\text{gcn}} = 3$ -layer graph, the node features of each design element are integrated with the topology information and type information of neighborhood elements, forming an element-level feature representation with enhanced spatial relationship:

$$H_{\text{gcn}} \in \mathbb{R}^{|V| \times D_{\text{gcn}}} \quad (44)$$

Among them:

$$D_{\text{gcn}} = 128 \quad (45)$$

This representation is further summarized into a global graph embedding vector by attention pooling [21],[22]:

$$c_{\text{graph}} \in \mathbb{R}^{D_{\text{graph}}} \quad (46)$$

Among them:

$$D_{\text{graph}} = 128 \quad (47)$$

Finally, c_{graph} and design constraint coding vector c_{cond} are concatenated into extended condition vectors:

$$c_{\text{ext}} = [c_{\text{cond}}; c_{\text{graph}}] \in \mathbb{R}^{D_{\text{ext}}} \quad (48)$$

Among them:

$$D_{\text{ext}} = 195 \quad (49)$$

The prior embedding of spatial relations enables the model to perceive the topological dependence between each design element and its surrounding elements when generating each design element, thus effectively suppressing the problems such as the conflict of adjacent relations, the violation of inclusion relations and the misalignment of axis relations. The task of the conditional embedding module is to transform the design constraints and spatial topology relations into numerical guidance signals and inject the denoising network to control the generation direction. [Table 3](#) summarizes the input source, dimension configuration and processing mode of each sub-module of the conditional embedding module. The output of the three sub-modules is spliced and fused to form the extended condition vector, which is used as the joint condition input of the denoising network.

Table 3. Configuration of the input and output dimensions of each sub-module of the conditional embedding module

Name of submodule	Source of input	Dimension of input	Dimension of output	Method of handling
Hard constraint coding	Land boundary/boundary withdrawal/fire protection	variable	$D_{\text{hard}} = 64$	Concatenation plus linear transformation
Soft constraint coding	View/enclosure/ratio index	$D_{\text{soft}} = 3$	$D_{\text{soft}} = 3$	Direct value taking
Spatial relation graph convolution	Design element diagram structure	$V \times D_{\text{node}}$	$D_{\text{graph}} = 128$	3-layer GCN+ pooling
Conditional extension fusion	The above three parts are spliced	$D_{\text{hard}} + D_{\text{soft}} + D_{\text{graph}}$	$D_{\text{ext}} = 195$	Fully connected mapping

In the hard constraint coding module, the land boundary mask is downsampled as a 16×16 binary matrix, the boundary-back minimum distance scalar value is spliced with the fire channel availability mark and mapped into 64-dimensional dense vector by linear transformation. The soft constraint coding module directly takes the original calculated values of the three continuous indexes of view permeability, spatial containment and area ratio deviation to form a three-dimensional vector and does not make nonlinear transformation to retain its physical interpretability. The spatial relationship graph

convolution module takes the attribute features of $|V|$ nodes in the design element graph structure as the input, aggregates neighborhood information through the three-layer graph convolution network and then summarizes it into a 128-dimensional global graph embedding vector by attention pooling, which explicitly encodes topological relationship information such as contiguity, inclusion and axis among design elements. The output of the three sub-modules is concatenated in sequence and then input into the conditional extension fusion module. Through full connection mapping, the 195-dimensional extension condition vector is generated, which unifies the feature expression scale of heterogeneous constraint information and prepares for the subsequent injection of each layer of the denoising network through adaptive layer normalization.

3.3 Denoising network backbone architecture

The denoising network is the core component of the diffusion model and is responsible for gradually predicting and removing noise from the noisy layout representation in the reverse process. In this study, U-Net is taken as the basic skeleton for targeted improvement, and dual-channel attention mechanism, collaborative injection strategy and progressive refinement generation strategy are introduced to make the denoising network have both high-fidelity reconstruction ability and structured layout generation ability.

The improved denoising network based on U-Net adopts encoderdecoder symmetry structure, including four downsampling stages and four upsampling stages, and high-resolution spatial details are transmitted between each stage through jump connection. The number of channels in each stage of the encoder is $[C, 2C, 4C, 8C]$, where the number of base channels C is set to 64, and the number of channels in each stage of the decoder is symmetrically decreasing. In each encoder stage and decoder stage, this study introduces a dual-channel attention module consisting of spatial attention and channel attention to enhance the selective attention of the network to important areas and key channels in the layout space [23],[24]. Let the input feature tensor be $X \in \mathbb{R}^{C \times H \times W}$. The spatial attention weight $A_{\text{spatial}} \in \mathbb{R}^{1 \times H \times W}$ is calculated as follows:

$$A_{\text{spatial}} = \text{Sigmoid} \left(\text{Conv}_{1 \times 1} \left([\text{MaxPool}(X) \text{AvgPool}(X)] \right) \right) \quad (50)$$

Among them, MaxPool and AvgPool take the maximum value and average value along the channel dimension respectively to obtain two $H \times W$ feature maps. After channel splicing, 1×1 convolution and Sigmoid activation are used to generate spatial attention weights. The channel attention weight $A_{\text{channel}} \in \mathbb{R}^{C \times 1 \times 1}$ is calculated as follows:

$$A_{\text{channel}} = \text{Sigmoid} \left(\text{MLP} \left([\text{MaxPool}_{\text{spatial}}(X) \text{AvgPool}_{\text{spatial}}(X)] \right) \right) \quad (51)$$

Among them, $\text{MaxPool}_{\text{spatial}}$ and $\text{AvgPool}_{\text{spatial}}$ take the maximum value and average value along the spatial dimension respectively to obtain two C -dimensional vectors, which are mapped back to C -dimension through multi-layer perception with a hidden layer after splicing, and then activated by Sigmoid.

The final output of the dual-channel attention module is the weighted sum of the original input modulated by the dual-channel weight:

$$X_{\text{attn}} = A_{\text{channel}} \odot (A_{\text{spatial}} \odot X) \quad (52)$$

Where $\odot$ indicates multiplication by element, and the weight of spatial attention is extended to

each channel through broadcast. The introduction of dual-channel attention enables the denoising network to dynamically emphasize the spatial regions and semantic channels strongly related to the quality of layout generation in the process of feature extraction and suppress the interference of irrelevant background features.

The cooperative injection mechanism of time-step embedding and conditional embedding determines which guidance signals the denoising network receives at different diffusion time steps. The forward procedure of the diffusion model encodes the time step t from 0 to T as a scalar that needs to be transformed into a dense vector suitable for the neural network. The time-step embedding vector $e_t \in \mathbb{R}^{D_{\text{time}}}$ is generated by combining sinusoidal position coding with multi-layer perception:

$$\text{PE}(t, 2k) = \sin\left(\frac{t}{10000^{\frac{2k}{D_{\text{time}}}}}\right) \quad (53)$$

$$\text{PE}(t, 2k + 1) = \cos\left(\frac{t}{10000^{\frac{2k}{D_{\text{time}}}}}\right) \quad (54)$$

$$e_t = \text{MLP}_{\text{time}}(\text{PE}(t, \cdot)) \quad (55)$$

Among them:

$$D_{\text{time}} = 64 \quad (56)$$

The range of k is $0, 1, \dots, \left(\frac{D_{\text{time}}}{2} - 1\right)$. Conditional embedding vector c_{ext} and timestep embedding vector e_t need to be injected in different ways at different locations in the network. In the encoder part, the time-step embeddings are injected through the adaptive layer normalization in the same way as the injection of the conditional vectors in Section 3.3. In the decoder part, in addition to the normalization injection through the adaptive layer, the conditional embedding is also channel-cascading with the encoder features through the feature modulation module at the jump junction, so that the decoding stage can simultaneously refer to the semantic guidance of the condition and the spatial details retained by the encoder.

The input feature $X_{\text{dec}}^{(l)}$ of the decoder layer l is output by the previous decoding layer, the corresponding encoder jumps connection features, and the projection features of conditional embedding are concatenated together and then obtained by convolution:

$$X_{\text{dec}}^{(l)} = \text{Conv}\left([\text{UpSample}\left(X_{\text{dec}}^{(l-1)}\right), X_{\text{skip}}^{(l)}, \text{Proj}(c_{\text{ext}})]\right) \quad (57)$$

Where Proj is a linear projection layer that maps c_{ext} from D_{ext} dimension to dimension that matches the number of channels in the current layer. The cooperative injection mechanism ensures the rhythm control of the denoising process guided by time information, and the compliance of the design constraints of the generation direction guided by conditional information. The two do not interfere with each other and complement each other in the whole network.

The denoising network is the core component of the diffusion model and is responsible for gradually predicting and removing noise from the noisy layout representation in the reverse process. In this study, U-Net is taken as the basic skeleton for targeted improvement, and dual-channel attention mechanism, collaborative injection strategy and progressive refinement generation strategy are

introduced [25],[26]. [Table 4](#) summarizes the input and output sizes, core parameter configurations and functional positioning of key modules of denoising network. Each module is connected step by step along the encoder-decoder path to form a complete computing link from global semantic abstraction to local detail reconstruction.

Table 4. Configuration parameters of key modules of denoising network

Name of module	Size of input	Dimension of output	Core parameter
Encoder Stage 1	256×256×3	128×128×64	Step 2 convolution
Encoder Stage 2	128×128×64	64×64×128	Step size 2 convolution + spatial attention
Encoder Stage 3	64×64×128	32×32×256	Step size 2 convolution + dual-way attention
Encoder Stage 4	32×32×256	16×16×512	Step size 2 convolution + dual-way attention
Decoder refinement module	16×16×512	256×256×3	Void convolution pyramid
Output projection layer	256×256×64	256×256×7	1 by 1 convolution

The encoder part consists of four downsampling stages, and the number of channels in each stage is multiplied in turn. The number of base channels is set to 64. Encoder Stage 1 halves the spatial resolution to 128 by 128 and expands the number of channels to 64 through convolution with step size of two to complete the preliminary extraction of low-level visual features such as edges and textures. Encoder Stage 2 introduces the spatial attention module on the basis of step 2 convolution, so that the network can focus on the spatial area carrying key design elements in the garden plane, and the output scale is 64 by 64 by 128. Encoder Stage 3 and Stage 4 are equipped with dual-channel attention modules, which pay attention to the importance of spatial location and channel semantics at the same time. Stage 4 outputs a global context feature map of 16 by 16 by 512, which carries the semantic abstraction of the overall garden land contour and spatial sequence structure.

The decoder part gradually recovers the spatial resolution through upsampling and jumps connection. The decoder refinement module takes the deepest 16 by 16 by 512 features of the encoder as the input, uses the cavity convolution pyramid to extract multiple receptive field features in parallel with the expansion rate of one, two, four and eight, and then outputs the refined features through attention fusion, which effectively improves the sharpness of the generated layout at the boundary of elements. The output projection layer is 1×1 convolution, and the number of channels is mapped from 64 to seven to generate the probability map of each design element category, which is used as the rasterized layout prediction of the final output of the denoising network. In the above configuration, the encoder expands the receptive field and improves the semantic abstraction level through gradual downsampling at each stage, and the decoder refinement module compensates for the loss of spatial detail caused by downsampling through multi-scale hole convolution. The two cooperate to realize the complete generation function from global structure control to local boundary refinement.

3.4 Model training strategy and loss function design

The complete generative model training involves the optimization of denoising diffusion process, the constraint of spatial relationship consistency and the guidance of design constraint compliance. In

this study, a composite loss function and a phased training strategy are designed to make each objective co-optimized. The training objective takes the theoretical framework of conditional denoising diffusion probability model as the starting point. In the forward stage of the diffusion process, the clean data x_0 is noise-added to x_t , and its conditional probability distribution is as follows:

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}; x_0(1 - \bar{\alpha}_t)I) \quad (58)$$

Among them:

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s \quad (59)$$

$$\alpha_s = 1 - \beta_s \quad (60)$$

Is the predefined noise scheduling parameter:

$$\beta_1 = 10^{-4}, \beta_T = 0.02, T = 1000 \quad (61)$$

$$\mathcal{N}(\cdot; \mu; \sigma^2) \quad (62)$$

Denote the Gaussian distribution with mean μ and variance σ^2 , and I is the identity matrix. The goal of the reverse denoising network is to predict the noise $\varepsilon \sim \mathcal{N}(0, I)$ applied to x_t under the given time step t and condition c_{ext} . The training loss of the standard diffusion model is the mean square error of noise prediction:

$$L_{\text{diff}} = \mathbb{E}_{t, x_0, \varepsilon} [\|\varepsilon - \varepsilon_\theta(x_t, t, c_{\text{ext}})\|_2^2] \quad (63)$$

Where ε_θ is the predicted value of noise by the denoising network, θ is, the trainable parameter of the network, and $\mathbb{E}_{t, x_0, \varepsilon}$ represents the joint expectation of time step t , clean data x_0 and standard Gaussian noise ε . This loss function enables the network to gradually learn the ability to recover clean layouts from noisy data by minimizing the Euclidean distance between the predicted noise and the true noise.

The composite loss function adds the spatial relationship consistency loss and the design constraint violation penalty on the basis of the standard diffusion loss, so that the model not only pays attention to the reconstruction accuracy at the pixel level but also pays more attention to the structural rationality and specification compliance at the layout level. The consistency loss of spatial relations L_{rel} directly supervises whether the topological relations between design elements in the generated layout are correct. For each pair of design elements (v_i, v_j) , the predicted relationship type is r_{ij}^{pred} and the true relationship type is r_{ij}^{gt} , then the relationship consistency loss is defined as:

$$L_{\text{rel}} = \frac{1}{|E|} \sum_{e_{ij} \in E} \text{CE}(r_{ij}^{\text{pred}}, r_{ij}^{\text{gt}}) \quad (64)$$

Where CE represents the cross-entropy loss and $|E|$ is the total number of edges in the graph. The relation prediction head is attached to the end of the decoder and is mapped by the classification layer from the output H_{gcn} of the spatial relation prior embedding module. The design constraint violation penalty term $L_{\text{constraint}}$ continues the definition of section 2.4 but unifies the hard and soft penalties into a single scalar loss term and controls its contribution proportion to the overall loss through the dynamic weight adjustment strategy. The complete form of the composite loss function L_{total} is:

$$L_{\text{total}} = L_{\text{diff}} + \lambda_{\text{rel}}L_{\text{rel}} + \lambda_{\text{constraint}}L_{\text{constraint}} \quad (65)$$

Where λ_{rel} and $\lambda_{\text{constraint}}$ are the weight coefficients to balance the contributions of various items, which are taken after preliminary debugging in this study:

$$\lambda_{\text{rel}} = 0.3, \lambda_{\text{constraint}} = 0.5 \quad (66)$$

In the early stage of training (the first 50 epochs), the model has not yet formed a reasonable layout generation ability, and it is difficult for the supervision signals of spatial relations and constraint compliance to provide effective gradients. Therefore, the warm-up strategy is set as follows:

$$\lambda_{\text{rel}} = \lambda_{\text{constraint}} = 0 \quad (67)$$

It increases linearly to the target value from the 50th to the 100th epoch and remains unchanged thereafter.

The phased training strategy further refines the optimization process of the model to solve the gradient conflict and convergence difficulties that are easy to occur in multi-objective joint training. The first stage is the pre-training of the constrained embedding module, which lasts for 30 epochs. In this stage, the backbone parameters of the denoising network are fixed, and only the parameters of the convolution part of the design constraint coding and spatial relation graph in the conditional embedding module are updated. The training objective is simplified as $L_{\text{constraint}}$, so that the conditional embedding module can quickly converge to the state that can effectively map the design constraints into meaningful feature vectors.

The second stage is the joint fine-tuning of the denoising network, which lasts 120 epochs. In this stage, all module parameters are trained openly, and the compound loss function L_{total} fully participates in the gradient back transmission. The optimizer uses AdamW, the learning rate is set to 10^{-4} and the cosine annealing scheduling strategy is used to gradually decay to 10^{-6} during the training process. The batch size is set to 8 and the gradient clipping norm is set to 1.0 to prevent gradient explosion. The loss value on the validation set is calculated at the end of each epoch. If the validation loss does not decrease for ten consecutive epochs, the early stop mechanism is triggered to terminate the training and save the best model weight.

4. MODEL EXPERIMENT VERIFICATION AND RESULT ANALYSIS

4.1 Experimental Setup

In terms of experimental environment configuration, all model training and reasoning are completed on a unified hardware platform. The compute node is configured with a dual-channel Intel Xeon Gold 6248R processor with a frequency of 3.0 GHz and a system memory of 256GB DDR4. The accelerated computing resources are four NVIDIA A100 Tensor Core Gpus, each with 80GB of video memory capacity, interconnected by NVLink for multi-card communication. PyTorch 2.1.0, CUDA 12.2, and cuDNN 8.9 are used as a deep learning framework. In the training process, multi-process parallel mode is adopted for data loading, and the number of prefetching threads is set to 8 to eliminate the impact of I/O bottleneck on training efficiency. The training hyperparameters are set as follows: The total number of steps of denoising diffusion is $T=1000$, the cosine scheduling strategy is used for noise scheduling, AdamW is used as the optimizer, the initial learning rate is 1×10^{-4} , the weight attenuation coefficient is 1×10^{-5} , the total number of training rounds is 200 epochs, the batch size is 8, and the

gradient accumulation step is 2 to achieve the effect of batch size 16. The learning rate was reduced to 1×10^{-6} in the middle of training (the 120th epoch) by cosine annealing with a gradient clipping threshold of 1.0. The total number of model parameters is about 147 million, the time of a single forward inference is about 2.3 seconds on a single A100 block, and the total time of generating a garden layout scheme (including 1000 denoising iterations) is about 38 seconds.

The selection of the comparison model covers the mainstream methods in the field of generative layout design to ensure the comprehensiveness and fairness of the comparison. [Table 5](#) summarizes the basic architecture characteristics and parameter scale of each comparison model. All baseline models train the same number of epochs on the same data set and adopt the hyperparameter configuration recommended by the original literature.

Table 5. compares the basic architectural features of the model with the parameter scale

Name of model	Normal form of generation	The core architecture	Conditional injection mode	Parameter scale (M)
LayoutGAN	Generative adversarial network	Graph convolution generator + discriminator	There are no explicit conditions	42.6
LayoutVAE	Variational autoencoder	RNN encoder-decoder	Sampling of hidden variables	28.3
Standard diffusion model	Diffusion probability model	U-Net	Time step only	86.7
House-GAN++	Generative adversarial network	Diagram + generator	Graph constrained embedding	55.1
Model of this study	Diffusion probability model	Multi-scale U-Net+ conditional embedding	Constraints + relational priors	147.2

LayoutGAN adopts the generative adversarial network framework to achieve vector generation of layout through graph convolution generator and discriminator. Its parameter scale is 42.6 million, which is the smallest among all comparison models, but it lacks explicit conditional injection mechanism. LayoutVAE is based on the variational autoencoder architecture and uses RNN encoder-decoder with hidden variable sampling for serialization layout generation, with a parameter size of 28.3 million. As the most direct comparison baseline, the standard diffusion model adopts DDPM original U-Net architecture with only time-step embedding as conditional input and a parameter scale of 86.7 million. Based on the generative adversarial network, House-GAN++ restricts the topological validity of the layout through a relationship graph with a parameter scale of 55.1 million. On the basis of the above baseline, this research model introduces multi-scale spatial feature encoder, conditional embedding module and spatial relation graph convolution prior. The parameter scale is 147.2 million, which is the largest, but the experimental part will prove that its performance gain far exceeds the computational cost caused by the increase of parameters.

4.2 Definition of quantitative evaluation indicators

The quantitative evaluation index comprehensively evaluates the model performance from three dimensions of layout rationality, generation diversity and design quality. Each dimension contains several fine-grained sub-indexes, and the mean and standard deviation of all indexes are calculated on the 30 samples of the test set.

The layout rationality dimension measures the effectiveness of the generation scheme in physical space and functional logic, including three sub-indexes: satisfaction rate of design constraints, spatial coverage rate and element integrity rate. The satisfaction rate of design constraints is further subdivided into hard constraint violation rate and soft constraint deviation degree. The hard constraint violation rate $R_{\text{violation}}$ is defined as the proportion of samples that violate at least one hard constraint in the generated layout scheme:

$$R_{\text{violation}} = \frac{1}{N_{\text{test}}} \sum_{m=1}^{N_{\text{test}}} I \left(v_{\text{hard}}^{(m)} = 1 \right) \quad (68)$$

Where $N_{\text{test}} = 30$ is the total number of samples in the test set, and $I(\cdot)$ is the indicator function, which is 1 when the condition in parentheses is true, otherwise it is 0; $v_{\text{hard}}^{(m)}$ is the hard constraint violation flag in the m -th generation scheme, as defined in Section 2.4. soft constraint deviation degree $D_{\text{soft,avg}}$ is defined as the average normalized deviation of each soft constraint index relative to the expected threshold:

$$D_{\text{soft,avg}} = \frac{1}{3N_{\text{test}}} \sum_{m=1}^{N_{\text{test}}} \left[\frac{\max \left(0, V_{\text{target}} - V_{\text{openness}}^{(m)} \right)}{V_{\text{target}}} + \frac{\max \left(0, C_{\text{target}} - C_{\text{enclosure}}^{(m)} \right)}{C_{\text{target}}} + D_{\text{ratio}}^{(m)} \right] \quad (69)$$

Where $V_{\text{openness}}^{(m)}$, $C_{\text{enclosure}}^{(m)}$, $D_{\text{ratio}}^{(m)}$ are the viewport permeability, spatial containment and area ratio deviation values of the m -th generation scheme, respectively. $V_{\text{target}} = 0.45$, $C_{\text{target}} = 0.55$ are the expected thresholds, and the lower the $D_{\text{soft,avg}}$ value, the higher the compliance degree of soft constraints. Spatial coverage C_{overlap} measures the ratio of the total area occupied by all design elements in the generated layout to the total area of garden land. The calculation formula of spatial coverage is as follows:

$$C_{\text{overlap}} = \frac{1}{N_{\text{test}}} \sum_{m=1}^{N_{\text{test}}} \frac{\sum_{i=1}^{|V_m|} \text{Area} \left(A_i^{(m)} \right)}{\text{Area} \left(\Omega_m \right)} \quad (70)$$

Where $|V_m|$ is the total number of design elements in the m -th generation scheme, $A_i^{(m)}$ is the Area occupied by the i -th element, $\text{Area}(\cdot)$ is the area function, Ω_m is the garden land boundary corresponding to the m -th scheme.

Element completeness rate R_{complete} measures the completeness degree of design elements in each functional partition in the generation scheme, which is defined as the proportion of the number of element categories actually generated to the number of expected element categories in the partition. The generative diversity dimension is used to evaluate whether the model can generate rich and diverse design schemes rather than the same and repeated output, including three sub-indicators: Fresher distance, category distribution entropy and number of patterns. Frechet Inception Distance (FID) was originally used to evaluate the distribution difference between the generated image and the real image. In this study, it was adapted to the distribution area, and the pre-trained autoencoder was used to project the raster layout into the feature space, and then the Frechet distance between the feature distribution of

the generated set and the real set was calculated:

$$\text{FID} = \|\mu_{\text{gen}} - \mu_{\text{real}}\|_2^2 + \text{Tr} \left(\Sigma_{\text{gen}} + \Sigma_{\text{real}} - 2(\Sigma_{\text{gen}}\Sigma_{\text{real}})^{1/2} \right) \quad (71)$$

Where $\mu_{\text{gen}}, \mu_{\text{real}}$ are the mean vectors of the generated set and the real set in the feature space respectively, $\Sigma_{\text{gen}}, \Sigma_{\text{real}}$ are the corresponding covariance matrices respectively, and $\text{Tr}(\cdot)$ is the trace operation of the matrix. A lower FID value indicates that the distribution of the generated layout is closer to the distribution of the true layout.

Category distribution entropy H_{cat} measures the evenness of the frequency distribution of various design elements in the generation scheme, and its calculation formula is:

$$H_{\text{cat}} = - \sum_{k=1}^C p_k \log_2(p_k) \quad (72)$$

Where p_k is the normalized value of the occurrence frequency of the k -th design element in the generated layout, and $C = 7$ is the total number of design element categories. High category distribution entropy indicates that the model can generate all kinds of design elements in a balanced manner without obvious category preference bias. The number of patterns N_{mode} uses the k-means clustering method to estimate the approximate number of different layout patterns in the generation set. The specific method is to cluster the feature vectors of all the generation schemes and determine the optimal number of clusters K by the contour coefficient. K is used as the estimate value of the number of patterns.

The quantitative indicators of design quality directly evaluate the advantages and disadvantages of the generated scheme in the professional dimension of landscape design, including spatial enclosure degree, visual permeability and deviation of the area ratio of functional zoning. These three indicators are consistent with the definition of soft constraints in section 2.4. The spatial enclosure degree $C_{\text{enclosure}}$ directly calculates the boundary of surface elements enclosed by adjacent elements according to the formula in section 2.4. The viewport permeability V_{openness} also follows the formula in section 2.4 Calculate the average visible distance of the observation point. Functional partition area ratio deviation D_{ratio} . The sum of squared deviations between the area proportions of the four types of zoning and the recommended proportions was calculated according to the formula in section 2.4.

Table 6 summarizes the calculation methods, theoretical value intervals and optimal value directions of all ten quantitative evaluation indicators, providing a unified quantitative basis for the subsequent comparison experiments of various models.

Table 6. Calculation method and value interval of each quantitative evaluation index

Dimension of dimension	Name of indicator	Method of calculation	Interval of value	Direction of merit
Rationality of layout	Hard constraint violation rate	Proportion of samples violating hard constraints	[0, 1]	↓ As low as possible
Rationality of layout	Soft constraint deviation degree	The normalized mean of the deviations from the constraint	[0, +∞)	↓ As low as possible
Rationality of layout	Coverage of space	Total factor area/total land area	[0, 1]	Center is better
Rationality of layout	Element integrity ratio	Actual number of categories/expected number of categories	[0, 1]	↑ The higher the better
Diversity of generation	Fresher distance	Feature distribution difference measure	[0, +∞)	↓ As low as possible
Diversity of generation	Category distribution entropy	The entropy of the class probability	[0, log ₂ 7]	↑ The higher the better
Diversity of generation	Number of modes	Cluster the optimal K value	[1, N_{test}]	↑ The higher the better
Quality of design	Degree of space enclosure	The boundary is enclosed in proportion	[0, 1]	↑ The higher the better
Quality of design	Field of view permeability	Average visible distance/feature scale	[0, 1]	↑ The higher the better
Quality of design	Deviation of area ratio	Sum of squared deviations of partition proportion	[0, +∞)	↓ As low as possible

The dimension of layout rationality contains four sub-indicators. The optimal values of the violation rate of hard constraints and the deviation degree of soft constraints are both the lower the better. The former reflects the model's compliance ability with the three mandatory norms of land use red line, boundary withdrawal requirement and fire passage, while the latter reflects the satisfaction degree of the three optimization conditions of view permeability, spatial containment and area ratio. The space coverage rate should be in the middle, and the ideal value should be between 70 and 80 percent to avoid inadequate space utilization or excessive density of elements. The higher the element integrity rate, the better, which measures the completeness of design elements within each functional zone.

The generative diversity dimension contains three sub-indicators. The lower the Fresher distance, the better, measures the distribution difference between the generated layout and the true layout in the feature space. The higher the category distribution entropy, the better measures the evenness of the distribution of the seven types of design elements in the frequency of use. The higher the number of patterns, the better. The optimal number of clusters is determined by K-means clustering of the feature vectors of the generated schemes to reflect the structural type of richness of the generated schemes.

The design quality quantification dimension contains three sub-indicators. The direction of the optimal value of spatial closeness and visual permeability is that the higher the better. The former quantifies the proportion of the boundary of surface elements enclosed by adjacent elements, and the

latter quantifies the ratio of the average visible distance of observation points in all directions to the garden characteristic scale. The lower the deviation of the area ratio, the better, and the sum of squared deviations between the actual area ratio and the recommended ratio of the four functional zones were calculated to reflect the equilibrium degree of the macro functional organization.

4.3 Generate layout quality comparison experiment

[Figure 1](#) shows the comparison of typical layout schemes generated by each model under the same garden land condition. A test sample with a rectangular shape (40 meters long and 30 meters wide) and an area of 1,200 square meters is selected as the display object.

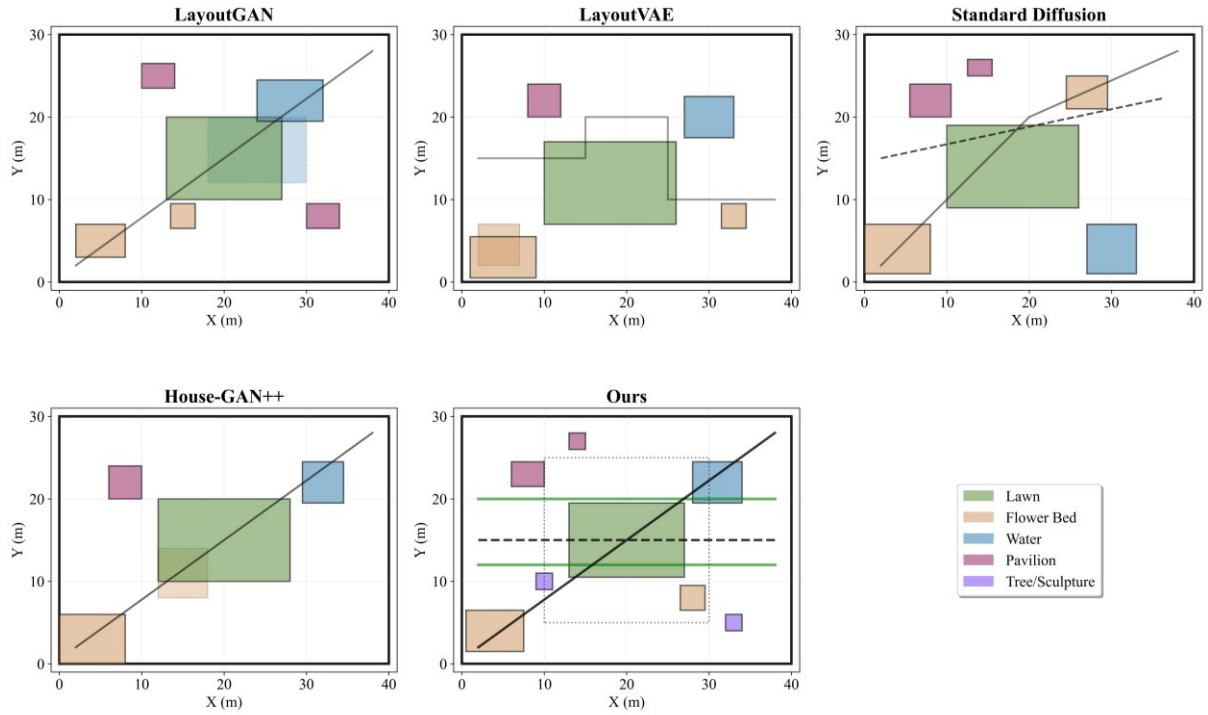


Figure 1. Comparison of sensory garden layout schemes generated by each model under the same land use conditions

The scheme generated by LayoutVAE has regular shape of single elements but loose spatial relationship, and the overall layout is characterized by fragmentation. The visual realism of the standard diffusion model generation scheme is better than the former two, but the main axis of the entrance area and the central viewing area is offset by about 12 degrees, which destroys the spatial sequence continuity. The topological relationship of the scheme generated by House-gan+ + is well maintained, but the element configuration is single, and the ornamental area lacks the combination of isolated planting trees and landscape sketches. The functional zoning boundaries of the scheme generated by the model in this study are clear and the area ratio is reasonable. The entrance area and the central viewing area form a clear axis alignment. The rest of the area is equipped with isolated trees and enclosed hedges to form a semi-private space. [Table 7](#) summarizes the complete scores of each model in all the ten quantitative evaluation indicators, which also covers the quantitative results of the three dimensions of layout rationality, generation diversity and design quality.

Table 7. Summary of all quantitative evaluation index scores of each model on the test set

Index of evaluation	LayoutGAN	LayoutVAE	Standard model	diffusion	House-GAN++	Model of this study
Hard constraint violation rate (%)	23.1	34.7	14.5		18.6	2.3
Deviation degree of soft constraint (%)	29.4	38.1	15.3		21.7	6.8
Space coverage (%)	72.4	65.8	78.2		74.9	76.3
Element integrity rate (%)	71.6	63.4	82.7		79.2	92.8
Fresher distance	21.6	28.3	14.2		17.8	8.7
Category distribution entropy	1.82	1.54	2.13		1.96	2.41
Number of modes	7	5	10		8	12
Space enclosure degree (%)	46.7	42.3	53.8		50.2	59.3
Visual field permeability (%)	43.2	39.8	46.5		44.7	49.7
Area ratio deviation ($\times 10^{-3}$)	8.9	12.4	6.7		9.3	5.2

The comparison of quantitative indicators shows that the research model has achieved the best results in all ten indicators. In the dimension of layout rationality, the violation rate of hard constraints is only 2.3 percent, which is about 84 percent lower than that of the sub-optimal standard diffusion model. The deviation degree of the soft constraint is 6.8 percent, which is better than 15.3 percent in the standard diffusion model. In the dimension of generative diversity, the Fresher distance is 8.7, the category distribution entropy is 2.41, and the number of patterns is 12, which are significantly better than all the comparison models. In the dimensions of design quality, the spatial enclosure degree is 59.3%, the visual field permeability is 49.7%, and the area ratio deviation is 0.052, which are all optimal, indicating that the model has superior generation diversity and constraint compliance ability while maintaining high design quality.

4.4 Ablation Experiments

Validation of the conditional embedding module was evaluated by removing the entire conditional embedding branch. In this ablation configuration, the denoising network only receives time-step embeddings as input, completely abandoning the guidance of design constraints and spatial relationship priors. [Figure 2](#) illustrates the degradation of the generated layout in terms of constraint compliance before and after the removal of conditional embeddings. After removing the conditional embeddedness, the violation rate of the hard constraints of the model increases from 2.3 percent to 43.6 percent, and a large number of generated schemes have problems such as design elements crossing the land boundary, insufficient boundary distance and lack of fire access. The results show that although the diffusion model verifythe guidance of conditional constraints can generate visually reasonable layout patterns, it does not have the ability to perceive and comply with the engineering design specifications, which fully verifies the key role of conditional embedding module as the transmission channel of design constraints.

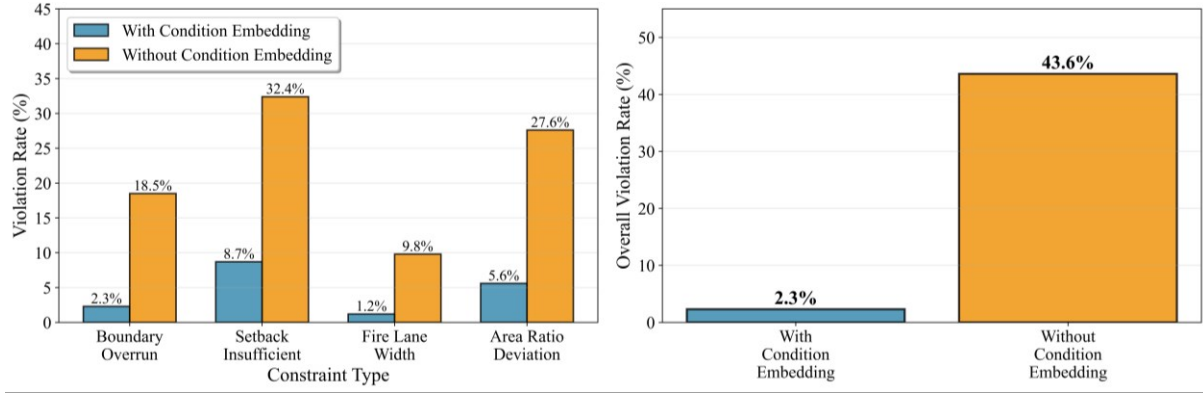


Figure 2. Distribution comparison of hard constraint violation after removing conditional embedding module (with/without embedding)

The contribution analysis of each scale of the multi-scale encoder removes the global scale coding branch and the local scale coding branch respectively for comparison experiment. After removing the global scale coding branch, the scale coordination of the layout generated by the model decreased significantly among the functional zones, specifically, the standard deviation of spatial coverage increased from 3.2% to 11.8%, and the deviation of area ratio increased from 0.052 to 0.0163. This shows that the lack of global information leads to the inability of the model to effectively control the relative area ratio of each partition. After removing the local scale coding branches, the blurring degree of the element boundary of the generated layout increases, the element integrity rate decreases from 92.8 percent to 784 percent, and the local combination pattern of elements (such as the collocation of various plants in the flower area) tends to be unitary. When the global and local coding branches are used together, all the indexes reach the optimal, indicating that the two scale coding branches capture complementary spatial information respectively, and the synergistic effect of the two branches is greater than the simple superposition of their own effects.

The gain of the spatial relationship prior module is verified by observing the change of the model in the topological relationship preservation ability after removing the graph convolution module. In the standard configuration, the spatial relationship prior module explicitly encodes the contiguity, inclusion and axis relationships between design elements through graph convolutional networks. After removing the module, the consistency of the spatial relationship in the generation scheme deteriorates significantly, and the prediction accuracy of the relationship decreases from 91.2 percent to 76.5 percent, among which the accuracy of the axis relationship decreases most sharply from 86.3 percent to 66.7 percent. **Figure 3** compares the distribution curve of the deviation angle of the axis of the generated scheme with/without spatial relationship prior. This result fully shows that the graph convolution module effectively restricts the deviation of spatial relations in the generation process by encoding topological prior knowledge.

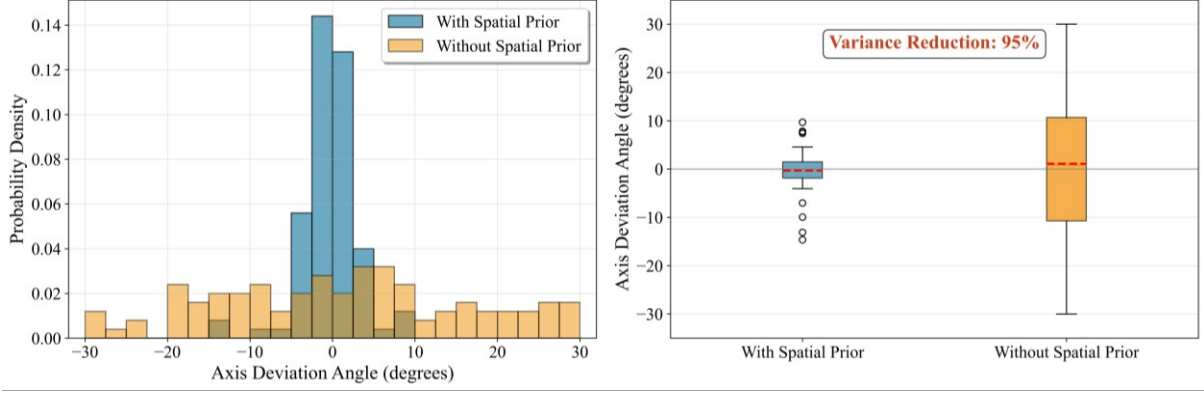


Figure 3. Influence of spatial relationship prior on axis deviation angle distribution (with/without prior comparison)

Contribution analysis of each loss function term evaluates the necessity of each loss term by removing the subterms in the composite loss function one by one. Three ablation variants were trained separately: removing spatial relationship consistency loss L_{rel} (denoted as w/o L_{rel}), removing design constraint penalty $L_{constraint}$ (denoted as w/o $L_{constraint}$), and removing both items at the same time retaining only L_{diff} (denoted as w/o Both). [Table 8](#) summarizes the performance of each ablation variant on key metrics. After removing L_{rel} , the violation rate of hard constraints increases slightly to 4.1 percent, but the accuracy of spatial relationship prediction decreases significantly to 76.9 percent, which confirms that L_{rel} has a specialized supervision role in topological relationship learning.

Table 8. Comparison of the performance of each variant in the ablation experiment on key evaluation indicators

Configuration of ablation	Hard constraint violation rate (%)	Deviation degree of soft constraint (%)	Relationship accuracy (%)	FID	Space enclosure degree (%)
The complete model	2.3	6.8	91.2	8.7	59.3
w/o L_{rel}	4.1	9.2	76.9	10.1	55.6
w/o $L_{constraint}$	23.7	12.5	88.5	11.4	51.2
w/o both	38.4	21.8	74.3	15.8	47.9

After removing $L_{constraint}$, the violation rate of hard constraints increases significantly to 23.7%, which is especially prominent in the aspect of boundary withdrawal constraints. After removing the two losses at the same time, the model performance degrades to the level close to the standard diffusion model, which verifies that each loss in the composite loss design has its own irreplaceable functional positioning and clear synergistic contribution.

4.5 Sensitivity analysis of design constraint strength

Changes in the strength of design constraints will affect the size of the feasible region and the sampling difficulty of the generation space. The backboundary distance was set as (1.0m), medium (2.0m) and strict (3.0m), and the area ratio tolerance was set as loose ($\pm 10\%$), medium ($\pm 5\%$) and strict

($\pm 3\%$), respectively. Thirty layout schemes were generated independently under each combination.

Figure 4 shows the trend curve of constraint satisfaction rate and design quality index from loose to strict constraints. The test results under different constraint strengths show regular patterns. The satisfaction rate of the hard constraints is 98.7 percent, and the deviation degree of the soft constraints is only 4.2 percent, but the spatial enclosure degree is 52.1 percent, which is lower than 59.3 percent under the medium constraints. Under the condition of strict constraints, the satisfaction rate of hard constraints decreased to 89.4 percent, the deviation degree of soft constraints increased to 12.8 percent, and the spatial enclosure degree reached 62.5 percent. As shown in **Figure 4**, the constraint satisfaction rate decreases with the increase of the intensity, but the slope is gentle, indicating that the model has good robustness to the change of the constraint intensity. The design quality index increases with the increase of intensity, but the growth rate decreases and tends to converge, indicating that the strict constraints guide the model to search for better design solutions, but it also gradually approaches the performance limit in the design space.

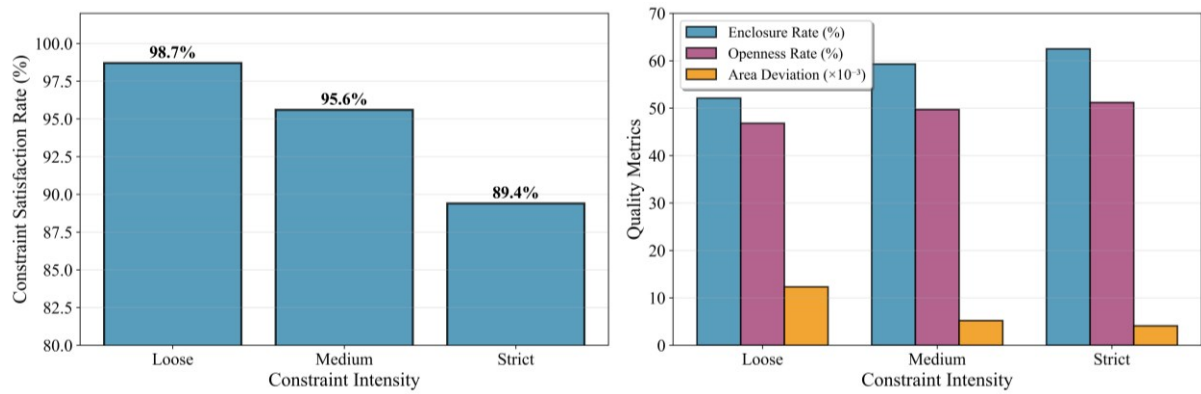


Figure 4. Variation trend of model constraint satisfaction rate and design quality under different constraint intensities

Constraint violation distribution analysis identifies the types of constraints that the model is most likely to violate during the generation process. In all 300 generated samples, a total of 184 violation incidents were classified by type: less than 47 percent of the boundary withdrawal distance, 26 percent of the area ratio deviation, 18 percent of the land red line crossing, and less than 9 percent of the fire channel width. Insufficient boundary withdrawal distance is the most important violation type, especially under the condition of strict boundary withdrawal, the model tends to lay elements close to the boundary area, resulting in a deviation of about 0.4m to 0.8m, indicating that the sensitivity mechanism of the model to the boundary distance needs to be strengthened.

The demonstration of multiple solutions verifies the ability of the model to generate multiple high-quality and different feasible solutions under the same design conditions. **Figure 5** shows four representative generation schemes under medium constraints of rectangular land use (40m $\times$ 30m), and each scheme shows systematic differences.

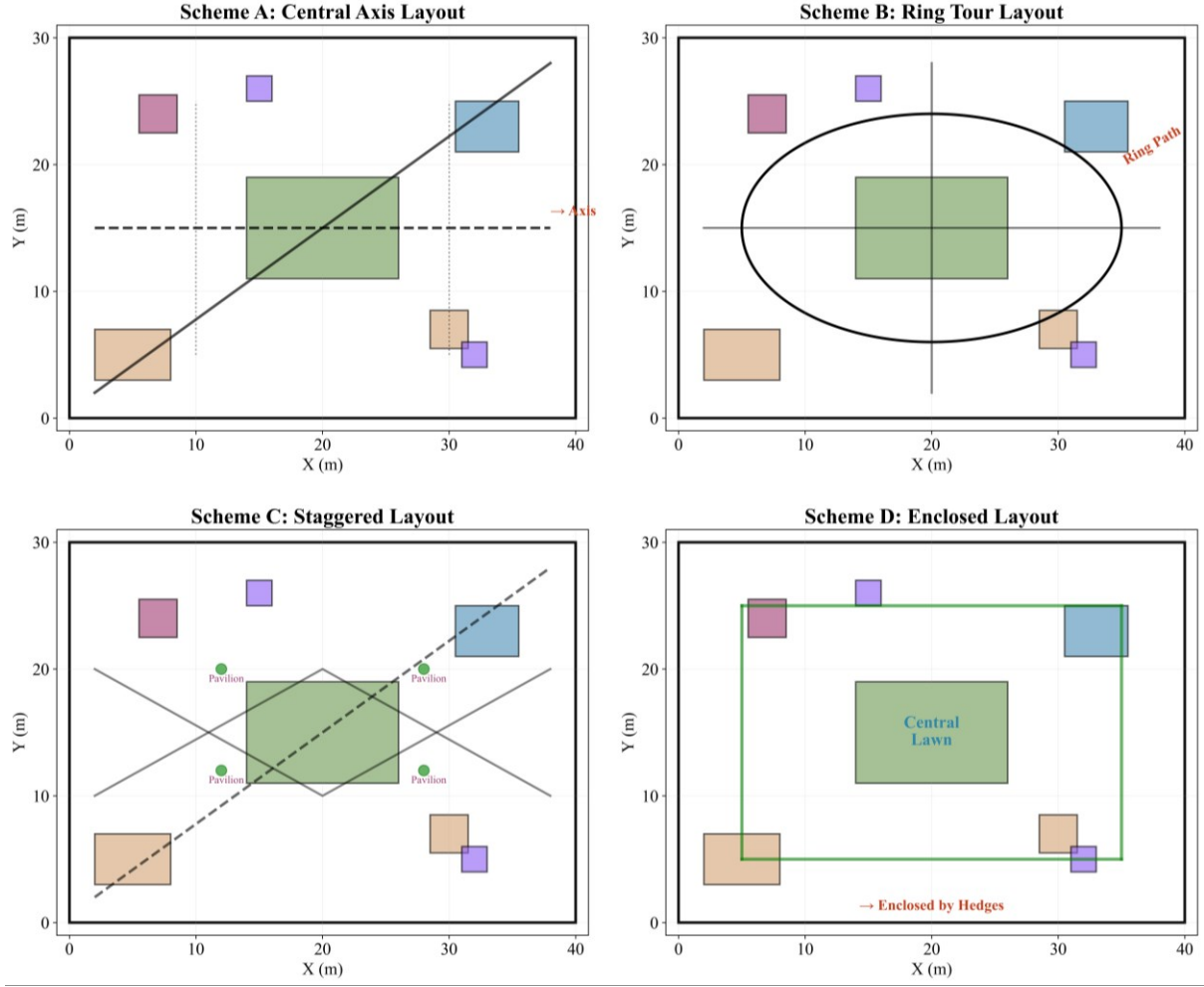


Figure 5. Four diversified layout schemes generated by the model under the same design conditions

Scheme A adopts the central axis through layout, the main axis from the entrance through the central water directly to the end of the rest pavilion; Plan B adopts a circular layout, with the land around the periphery of the park road and the central layout of each functional area; Scheme C adopts a scattered layout, with each functional area unfolding in turn along the diagonal direction; Plan D adopts an enclosed layout, with a large lawn set aside in the center and various ornamental elements arranged around it. The spatial coverage rate of the four schemes is between 74 % and 79 %, the violation rate of hard constraints is zero, the deviation degree of soft constraints is between 51 % and 83 %, and the spatial enclosure degree is between 57 % and 62 %, which all meet the design requirements and the quality difference is not significant.

4.6 Model generalization ability verification

Generalization ability verification evaluates the generality of the model from three perspectives: cross-shape generalization, cross-style generalization and small-sample adaptability.

The cross-garden shape generalization test directly applies the trained model to the irregular land where the training set does not appear, such as trapezoids, circles, triangles and irregular polygons. [Figure 6](#) shows the generated results of the model on five different shapes of land. The quantitative

results show that the violation rates of hard constraints on trapezoidal and circular land are 4.0% and 3.3% respectively, and the FID is 10.2% and 9.8% respectively. The violation rate of the hard constraints on the triangle and irregular polygon land increased to 12.0 and 10.7 percent, and the deviation of the area ratio increased significantly to 0.1558 and 0.1224. As shown in [Figure 6](#), the model can still identify boundaries and reasonably arrange elements on special-shaped land, but the phenomenon of element crowding or partial white space occurs in acute corners and narrow extension areas, indicating that the generalization ability of the model has been expanded beyond the training distribution but there is still room for improvement.

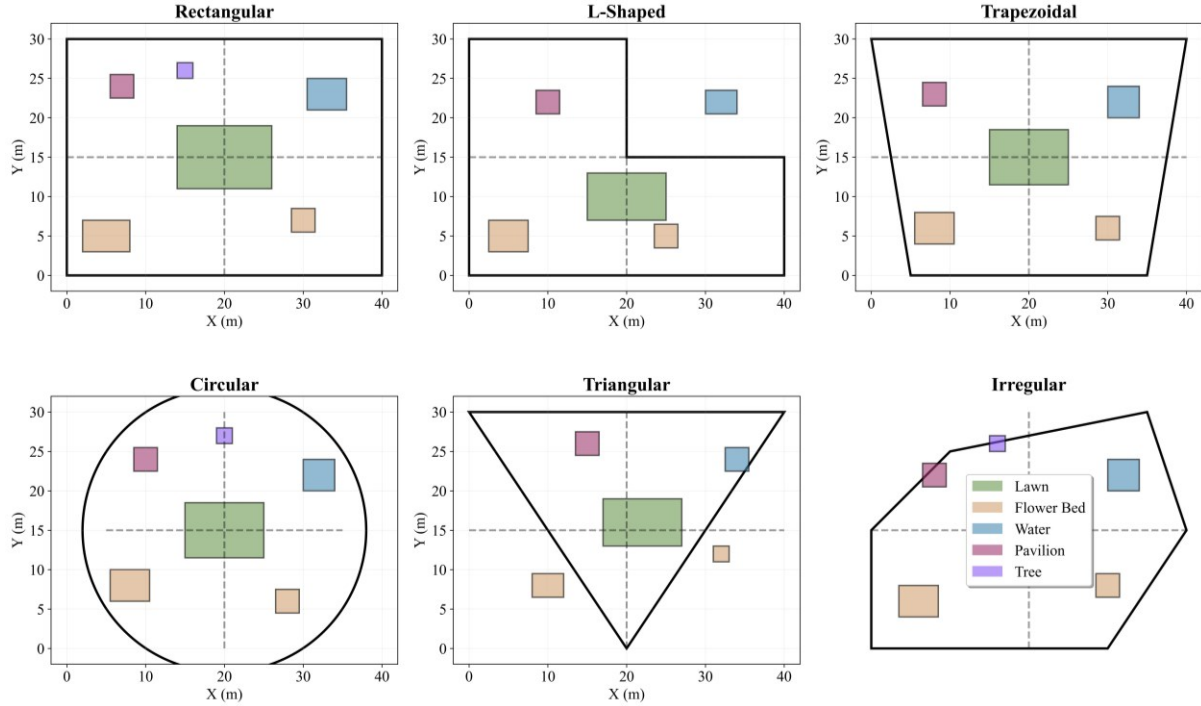


Figure 6. Generation results of cross-shape generalization of the model on five different shaped lands

Cross-design style generalization tests are generated according to three style labels: natural style, regular style and hybrid style. The natural style has the highest visual permeability (52.4 per cent) and moderate enclosure (56.7 per cent); Under the regular style, the axis relationship is the strictest (deviation angle standard deviation of 2.1 points) and the enclosure degree is the highest (63.8 percent). The hybrid is somewhere in between. The three styles show clear and discernible differences in visual features, and the satisfaction rates of all types of constraints are higher than 93 percent, indicating that the style coding in the conditional embedding module effectively guides the stylized expression of the generation direction.

Small sample scenario evaluation selects a subset of 10 to 100 percent from the full training set to train independent models. [Figure 7](#) shows the performance curve of FID and hard constraint violation rate as a function of training set size. When the number of training samples is 14, the FID is 25.3 and the violation rate of hard constraints is 18.7 percent. When the number of samples is increased to 70, the FID drops sharply to 12.5 percent, and the violation rate drops to 6.2 percent. The performance improvement flattens out after more than one hundred samples. As shown in [Figure 7](#), the inflection point of the curve is about 80 to 100 samples, indicating that the research model can obtain the generation ability with practical value after the training sample reaches this scale. This data efficiency

has practical significance for the limited high-quality case resources in the design field.

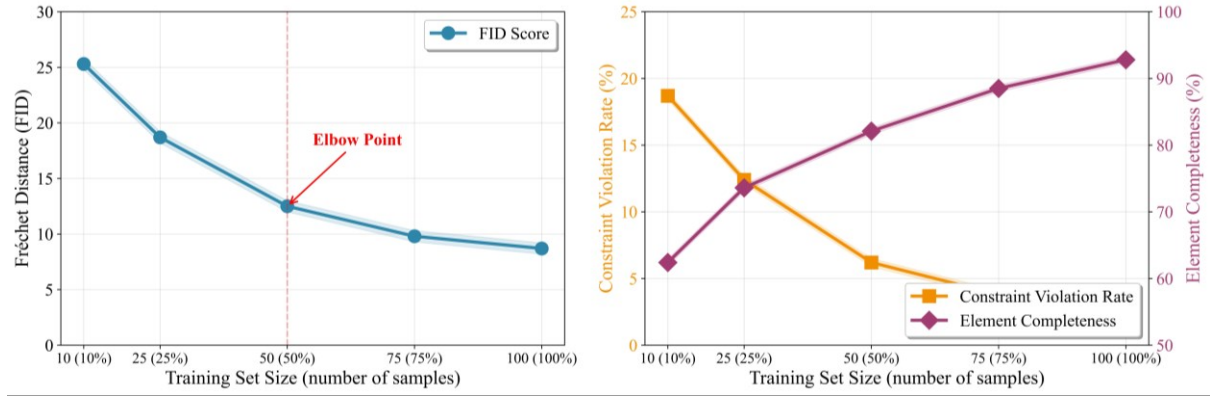


Figure 7. Performance change curve of model generation quality under different training set sizes

Based on the quantitative and qualitative results of all the experiments, the model proposed in this study shows good performance in terms of generation quality, constraint compliance, diversity, generalization ability and data efficiency, which verifies the effectiveness and practicality of the model architecture design.

5.DISCUSSION

Focusing on the generative layout problem of sensory garden environment design, this study constructs a deep learning model integrating design grammar rules and conditional constraints and makes a series of progress in data set construction, model architecture design and experimental verification. However, any research work has its inherent limitations and areas that can be improved.

In terms of methodological applicability, this study adopts denoising diffusion model as the basic generation framework. Although this paradigm is superior to generative adversative network and variational autoencoder in terms of generation quality and training stability, its iterative denoising generation mechanism determines that hundreds to thousands of forward propagations are required in the inference stage, and a single generation takes about 38 seconds. It is much higher than the second or even millisecond generation speed of generative adversarial network. Although real-time interaction is not considered as the primary optimization objective in the model design in this study, this computational overhead does limit the applicability of the model as a real-time aid tool in design practice. In addition, the high-dimensional feature space of the diffusion model and the raster input format make it have a certain mismatch in the processing of vectorized design elements, and the raster layout of the model output needs to be post-processed before it can be converted into vectorized design drawings, which may introduce additional accuracy loss. Future research can explore the potential diffusion model framework and complete layout generation in low-dimensional hidden space to significantly reduce reasoning costs and simplify the vectorization post-processing process.

In terms of the model performance boundary, although the experimental results prove the effectiveness of the model under standard test conditions, the problem of insufficient backboundary distance exposed in the ablation experiment deserves further discussion. The systematic characteristics of this deviation show that even if the hard constraint penalty term is adopted, the model still tends to arrange the elements close to the land under the strict boundary distance constraint. This point is further

confirmed in the cross-shape generalization experiment. When the model deals with the land use of triangles and irregular polygons, the factor crowding phenomenon in the acute corner area is the specific manifestation of the insufficient perception of boundary distance. This shows that the current constraint embedding method is still dominated by punitive guidance and does not yet have the ability to accurately control the spatial distance.

In terms of data dependence, the data set constructed in this study contains 200 valid samples, which are at a medium level in current generative design research driven by deep learning. Although the training scale is expanded to eight times through the data enhancement strategy, and the experiment shows that the practical generation ability can be obtained with 80 samples, the source bias of the data set cannot be ignored. Design competition schemes account for a relatively high proportion in the data set, with the average number of elements reaching 27 points, which is significantly higher than the 18.4 points of physical gardens built. In addition, rectangular and L-shaped land are dominant in the data set, which explains the phenomenon of model performance decline in triangular and irregular polygon land. Subsequent research can introduce transfer learning strategy, use large-scale general landscape layout data to pre-train the model, and then fine-tune the sensory garden special data, so as to alleviate the impact of data scarcity and distribution bias.

In terms of practical application suitability, this research model only supports two-dimensional plane layout generation at present and has not been extended to three-dimensional spatial form design. The relationship between terrain relief, vegetation vertical hierarchy and line of sight height difference in the sensory garden is an important factor affecting spatial experience, and the coupling relationship between two-dimensional layout and three-dimensional form has not been fully modeled in this study. In addition, although the layout scheme generated by the model meets the formal design constraints, it has not yet been optimized in linkage with engineering technical parameters such as sunshine analysis, drainage design and plant growth space requirements, which means that the generated scheme still needs to go through the secondary deepening of engineers before it can enter the construction drawing stage. At this stage, the model should be positioned as an auxiliary tool for scheme generation in the conceptual design stage, and its output results should be used as the reference starting point for designers rather than the final deliverable.

6. CONCLUSION

In this study, a generative layout design model driven by deep learning is proposed to solve the problems that traditional methods rely on manual experience, have low design efficiency and lack of scheme diversity in sensory garden environment design. This model takes the denoising diffusion probability model as the basic framework and realizes the end-to-end generation from the input of land use conditions and design constraints to the output of the complete garden layout plan through three core innovations: multi-scale spatial feature encoder, design constraint embedding module and spatial relation graph convolution prior. The model has completed training and system verification on the constructed data set, and the conclusions can be summarized as follows.

At the methodological level, this study verifies the feasibility and effectiveness of embedding diffusion models in the task of landscape layout generation in combination with design grammar rules. Through the hierarchical coding of global scale and local scale and cross-scale attention fusion, the multi-scale spatial feature encoder enables the model to perceive the overall spatial skeleton of the garden and the combination details of elements inside each functional zone at the same time, which

provides a solid feature foundation for the generation quality. The conditional embedding module encodes hard and soft constraints into differentiable conditional vectors and uses adaptive layer normalization to inject into each layer of the denoising network, so that the guidance signal of the design constraints runs through the whole generation process, rather than only performing posterior screening in the output stage. The spatial relationship graph convolution module makes up for the deficiency of the standard diffusion model in perceiving the topology structure by explicitly coding the neighborhood aggregation of the contiguity, inclusion and axis relationship between design elements. The synergy of the three improvements enables the model to endogenously comply with design specifications and grammar rules while maintaining visual realism.

At the level of experimental verification, the systematic comparison of ten quantitative evaluation indicators proves the superiority of the research model. In the dimension of layout rationality, the violation rate of hard constraints is as low as 2.3 percent, which is about 84 percent lower than that of 14.5 percent in the standard diffusion model. In the dimension of generation diversity, the Fresher distance reaches 8.7 and the category distribution entropy reaches 2.41, both of which are significantly better than the four baseline methods. In the dimension of design quality, the spatial enclosure degree reaches 59.3 percent, the deviation of area ratio is as low as zero zero fifty-two, and the comprehensive index is comprehensively leading. The ablation experiment further quantifies the independent contribution of each component, and the removal of conditional embedding module leads to the violation rate of hard constraints from 2.3 percent to 43.6 percent, and the removal of spatial relationship prior leads to an increase of about 78 percent in the variance of axis deviation angle, which fully verifies the necessity of various innovative mechanisms. The constraint sensitivity analysis shows that the model can maintain stable generation performance under different constraint intensities, and the parallel generation of four different styles of schemes under the same design condition shows the practical application value of the model. Cross-shape generalization and cross-style generalization experiments show that the model has the ability to extend to unfamiliar design conditions, and the small sample experiments reveal that 80 training samples can support the model to achieve practical level of data efficiency.

Taken together, the main contribution of this study lies in three aspects. Firstly, a coding system of design elements and a data set construction method for sensory garden environmental design are proposed, and a mathematical definition and a differentiable expression framework for design constraints are established, which provides a reusable data basis and a formal paradigm for problems for subsequent research. Secondly, a conditional diffusion generative model architecture integrating spatial relationship priors and design constraints is constructed, which provides methodological reference for generative design research in the field of landscape design. Thirdly, through systematic experimental verification and discussion, the capability boundary and limitation of the current model are clarified, and the key improvement directions such as boundary distance perception, cross-shape generalization and 3D expansion are pointed out.

Looking to the future, follow-up research can be advanced from three directions. In terms of model efficiency, the potential diffusion model framework or distillation acceleration technology is explored to compress the reasoning time from tens of seconds to seconds, providing the possibility of real-time human-machine collaborative design. In terms of function expansion, the generation of two-dimensional plane layout is extended to three-dimensional spatial form design, and dimensions such as terrain, vertical hierarchy of vegetation and line of sight analysis are introduced to construct a more complete digital twin generation framework for sensory garden. In terms of application integration, the model output is coupled with the physical environment simulation tool, and the performance indicators

such as sunshine, wind environment and drainage are fed back in real time during the generation process, so as to realize the performance-driven generative design closed loop. In addition, with the continuous accumulation of high-quality design data and the progress of cross-modal generation technology, it is a research direction with more long-term value to combine natural language design intention description with layout generation to build an intelligent design assistant with human-computer dialogue as the interactive interface.

Abbreviations

GAN, Generative Adversarial Network;
VAE, Variational Autoencoder;
CBR, Case-Based Reasoning;
GCN, Graph Convolutional Network;
RNN, Recurrent Neural Network;
MLP, Multi-Layer Perceptron;
DDPM, Denoising Diffusion Probabilistic Model;
FID, Fréchet Inception Distance;
IFLA, International Federation of Landscape Architects;
CHSLA, Chinese Society of Landscape Architecture;
RoI, Region of Interest;
AdamW, Adam with Weight Decay;
GPU, Graphics Processing Unit;
CUDA, Compute Unified Device Architecture;
cuDNN, CUDA Deep Neural Network library.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical

approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **J.Z.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **J.Z.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

- [1] He, M., Wang, Y., Wang, W. J., & Xie, Z. (2022). Therapeutic plant landscape design of urban forest parks based on the Five Senses Theory: A case study of Stanley Park in Canada. *International Journal of Geoheritage and Parks*, 10(1), 97-112. <https://doi.org/10.1016/j.ijgeop.2022.02.004>
- [2] Ren, Q., Weng, Y., & Hu, Z. (2025). A study of space creation for healing landscape design in the post-epidemic era. *Frontiers in Psychology*, 16, 1618451. <https://doi.org/10.3389/fpsyg.2025.1618451>
- [3] Shi, Y., Liu, S., Huang, X., & Sheng, Q. (2025). Roaming through perished gardens: the application of space syntax in the reconstruction of the 17th-century Zhi Garden. *Humanities and Social Sciences Communications*, 12(1), 1508. <https://doi.org/10.1057/s41599-025-05789-6>
- [4] Yan, A., & Cheng, Z. (2024). A review of the development and future challenges of case-based reasoning. *Applied Sciences*, 14(16), 7130. <https://doi.org/10.3390/app14167130>
- [5] Zheng, X., Huang, Y., Liu, Y., Zhang, Z., Li, Y., & Yan, H. (2025). Financing mode and scheme decision support for large urban rail transit projects: a revised case-based reasoning approach. *Engineering, Construction and Architectural Management*, 32(12), 8494-8523. <https://doi.org/10.1108/ECAM-03-2023-0202>
- [6] Wang, Y., Zhou, Z., Tan, X., Pan, Y., Yuan, J., Qiu, Z., & Liu, C. (2024). Unveiling the potential of progressive training diffusion model for defect image generation and recognition in industrial processes. *Neurocomputing*, 592, 127837. <https://doi.org/10.1016/j.neucom.2024.127837>
- [7] Singh, A., Chatta, N. K., Vagula, Y., Ehtesham, A., Kumar, S., & Talaei Khoei, T. (2026). A Taxonomy of Generative Models with a Focus on Diffusion Models and Denoising Techniques. *Electronics*, 15(6), 1293. <https://doi.org/10.3390/electronics15061293>
- [8] Liang, L., Miao, B., Wang, X., Akhtar, N., Vice, J., & Mian, A. (2026). CymbaDiff: Structured Spatial Diffusion for Sketch-based 3D Semantic Urban Scene Generation. *Advances in Neural Information Processing Systems*, 38, 81114-81138. <https://doi.org/10.48550/arXiv.2510.13245>
- [9] Zhou, Y., Leng, H., Meng, S., Wu, H., & Zhang, Z. (2024). StructDiffusion: End-to-end intelligent shear wall structure layout generation and analysis using diffusion model. *Engineering Structures*, 309, 118068. <https://doi.org/10.1016/j.engstruct.2024.118068>
- [10] Zeng, P., Yin, J., Huang, Y., Zhong, J., & Lu, S. (2025, October). AI-based generation and optimization of energy-efficient residential layouts controlled by contour and room number. In *Building Simulation* (Vol. 18, No. 10, pp. 2777-2805). Beijing: Tsinghua University Press. <https://doi.org/10.1007/s12273-025-1337-4>
- [11] Li, M. R., Cao, Y., & Li, G. W. (2023). An approach to developing and protecting linear heritage tourism: The construction of cultural heritage corridor of traditional villages in Mentougou District using GIS. *International Journal of Geoheritage and Parks*, 11(4), 607-623. <https://doi.org/10.1016/j.ijgeop.2023.11.002>
- [12] Meng, L., Zhang, B., & Cao, L. (2026). Establishing linear cultural heritage corridors by integrating cultural and ecological values: A case study of the Jinzhong section of the Great Tea Road. *Land*, 15(2), 293. <https://doi.org/10.3390/land15020293>
- [13] Guo, X., Fu, S., & Zhu, D. (2025). Aesthetic quality evaluation of packaging design with graph

neural networks and composition features. *Scientific Reports*, 15(1), 36046. <https://doi.org/10.1038/s41598-025-20046-1>

[14] Yuan, Y., Liu, F., Yang, G., & Wang, M. (2026). A review of representation and similarity measurement methods for geospatial scenes. *International Journal of Geographical Information Science*, 40(5), 1519-1546. <https://doi.org/10.1080/13658816.2025.2563697>

[15] Karimi, K. (2023). The configurational structures of social spaces: Space syntax and urban morphology in the context of analytical, evidence-based design. *Land*, 12(11), 2084. <https://doi.org/10.3390/land12112084>

[16] Waldner, F., Diakogiannis, F. I., Batchelor, K., Ciccotosto-Camp, M., Cooper-Williams, E., Herrmann, C., ... & Toovey, A. (2021). Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images. *Remote sensing*, 13(11), 2197. <https://doi.org/10.3390/rs13112197>

[17] Shu, D., Zhang, Z., Wan, F., Ru, W., Yang, B., Zhang, Y., ... & Chen, X. (2025). SatViT-Seg: A transformer-only lightweight semantic segmentation model for real-time land cover mapping of high-resolution remote sensing imagery on satellites. *Remote Sensing*, 18(1), 1. <https://doi.org/10.3390/rs18010001>

[18] Wang, Y., Zhang, W., Chen, W., & Chen, C. (2024). BSDSNet: Dual-stream feature extraction network based on segment anything model for synthetic aperture radar land cover classification. *Remote Sensing*, 16(7), 1150. <https://doi.org/10.3390/rs16071150>

[19] Yu, Z., Li, Y., Xiao, J., Zhou, H., & Lin, B. (2025). Physical embedding on building surface spatial relationship shows better performance on graph-based daylight prediction. *Building and Environment*, 114141. <https://doi.org/10.1016/j.buildenv.2025.114141>

[20] Han, J., Lu, X. Z., & Lin, J. R. (2025). Pretrained graph neural network for embedding semantic, spatial, and topological data in building information models. *Computer-Aided Civil and Infrastructure Engineering*, 40(26), 4607-4631. <https://doi.org/10.1111/mice.70073>

[21] Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph attention networks: a comprehensive review of methods and applications. *Future Internet*, 16(9), 318. <https://doi.org/10.3390/fi16090318>

[22] Li, Z. P., Wang, S. G., Zhang, Q. H., Pan, Y. J., Xiao, N. A., Guo, J. Y., ... & Huang, D. S. (2024). Graph pooling for graph-level representation learning: a survey. *Artificial Intelligence Review*, 58(2), 45. <https://doi.org/10.1007/s10462-024-10949-2>

[23] Ma, W., Jiang, Q., Wang, Q., Yu, D., Huang, Y., He, B., & Jin, X. (2025). Ycsc-unet: A y-shaped composite spatial channel network based on u-net for breast lesion ultrasound image segmentation. *Neurocomputing*, 131865. <https://doi.org/10.1016/j.neucom.2025.131865>

[24] Yang, F., & Wang, B. (2024). Dual Channel-Spatial Self-Attention Transformer and CNN synergy network for 3D medical image segmentation. *Applied Soft Computing*, 167, 112255. <https://doi.org/10.1016/j.asoc.2024.112255>

[25] Chen, Y., Huang, Q., Geng, M., Wang, Z., & Han, Y. (2025). A systematic review on cell nucleus instance segmentation. *IET Image Processing*, 19(1), e70129. <https://doi.org/10.1049/ipr2.70129>

[26] Luo, J. (2026). Design and Processing of Benevolent Image Symbols Based on Computer Graphics and Visual Communication. *International Journal of Image and Graphics*, 2750105.

<https://doi.org/10.1142/s0219467827501051>

APA Citation

Zhang, J. (2026). Deep Learning-Driven Generative Layout Design Model for Sensory Garden Environment Design. *Journal of Discovery Core*, 1(2), 1-37. <https://doi.org/10.67541/jdc2608>