

Article

# SceneDiffusion: Scene Generation Model Embedded with Spatial Constraints

Shanshan Yu <sup>1</sup>, Jiaxin Zhu <sup>1,2</sup>, Jiaqi Li <sup>1</sup>, Xunchun Li <sup>3</sup>, Kai Wang <sup>4</sup>, Jian Tu <sup>4</sup> and Danhuai Guo <sup>1,\*</sup> 

<sup>1</sup> School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; yuss@buct.edu.cn (S.Y.); zhujiabin1126@foxmail.com (J.Z.); lijiaqi@buct.edu.cn (J.L.)

<sup>2</sup> Beijing Sunwise Space Technology Ltd., Beijing 100004, China

<sup>3</sup> Academy of Broadcasting Science, National Radio and Television Administration of China, Beijing 100866, China; lixunchun@abs.ac.cn

<sup>4</sup> Beijing Institute of Aerospace Long March Vehicles, Beijing 100076, China; wkfing@163.com (K.W.); tj2779@163.com (J.T.)

\* Correspondence: gdh@buct.edu.cn

## Abstract

Spatial scenes, as fundamental units of geospatial cognition, encompass rich objects and spatial relationships, and their generation techniques hold significant application value in disaster simulation and emergency drills, delayed spatial reconstruction and analysis, and other fields. However, existing studies still face limitations in modeling complex spatial relationships during scene generation, leading to insufficient semantic consistency and geographical accuracy. The advancement of Geospatial Artificial Intelligence (GeoAI) offers a new technical pathway for the intelligent modeling of spatial scenes. Against this backdrop, we propose SceneDiffusion, a scene generation model embedded with spatial constraints, and construct a geospatial scene dataset incorporating spatial relationship descriptions and geographic semantics, aiming to enhance the understanding and modeling capabilities of GeoAI models for spatial information. Specifically, SceneDiffusion employs a spatial scene representation framework to uniformly characterize objects and their topological, directional, and distance relationships, enhances the interactive modeling of objects and relationships through a Spatial relationship Attention-aware Graph (SAG) module, and finally generates high-quality scene images conforming to geographic semantics using a Layout information-guided Conditional Diffusion (LCD) module. Both qualitative and quantitative experiments demonstrate the superiority of SceneDiffusion, achieving a 56.6% reduction in FID and a 35.3% improvement in SSIM compared to baseline methods. Ablation studies confirm the importance of multi-relational modeling with attention mechanisms. By generating scenes that satisfy spatial distribution constraints, this work provides technical support for applications such as emergency scene simulation and virtual scene construction, while also offering insights for theoretical research and methodological innovation in GeoAI.

**Keywords:** GeoAI; scene generation; spatial constraints; diffusion models; spatial relationship modeling; attention mechanisms



Academic Editors: Wolfgang Kainz, Hartwig H. Hochmair, Hao Li and Levente Juhász

Received: 23 April 2025

Revised: 22 June 2025

Accepted: 26 June 2025

Published: 27 June 2025

**Citation:** Yu, S.; Zhu, J.; Li, J.; Li, X.; Wang, K.; Tu, J.; Guo, D.

SceneDiffusion: Scene Generation Model Embedded with Spatial Constraints. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 250. <https://doi.org/10.3390/ijgi14070250>

**Copyright:** © 2025 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The groundbreaking advancements in artificial intelligence (AI) have brought new opportunities for the development of geospatial research. As a product of the deep integration between geospatial science and AI technology, Geospatial Artificial Intelligence

(GeoAI) is profoundly transforming how humans perceive and manage the Earth's physical environment by researching and developing machine spatial intelligence [1]. Scene generation stands as one of the core topics in GeoAI, aiming to bridge spatial cognition between maps and textual modalities. In this process, the spatial scene, serving as both the object and medium of spatial cognition, provides a fundamental framework for humans to comprehend and describe their surrounding environment. It encompasses not only physical entities (e.g., buildings, roads, water bodies, and vegetation) but also complex spatial relationships among them (e.g., topological structures, directional references, and distance metrics). By constructing visual scenes that incorporate spatial objects and their relationships, we can reveal distribution patterns of geographic entities, thereby deepening our understanding of complex geospatial systems.

Scene generation technology has broad applications across multiple domains, with growing research and practical demands, particularly under the rapid development of AI. For instance, in public safety, emergency drills are crucial for enhancing response capabilities, where their relevance and effectiveness directly impact training outcomes and real-world applicability. Simulating disaster scenarios helps participants familiarize themselves with procedures, improve skills, and optimize contingency plans. However, acquiring real-world map data is often hindered by security or privacy concerns, while traditional manual design methods are costly and lack diversity [2,3]. Generative AI can construct scenes tailored to practical needs, offering a controllable and efficient solution for emergency drills. Similarly, in fields such as ancient text studies or crime scene documentation, some spatial scenes exist only as textual descriptions, necessitating an automated technique to generate geosemantically compliant visual scenes from text to support further spatial analysis.

Despite the pressing demand, the complexity of spatial scenes poses challenges for GeoAI. Existing methods exhibit limitations in modeling complex spatial relationships, leading to insufficient semantic consistency and geographical accuracy in generated scenes. To address this, our study explores efficient mechanisms for embedding spatial information in deep learning models, developing a generation model that explicitly encodes spatial relationship constraints to achieve deeper scene understanding and advance intelligent applications.

Common representations of spatial scenes include natural language descriptions, sketch-based depictions, and scene graphs [4–6]. Among these, the scene graph, a structured representation using nodes and edges to describe objects and their relationships, effectively supports complex spatial analysis. Leveraging natural language processing (e.g., relation extraction) and computer vision techniques (e.g., object detection), textual and sketch-based representations can be uniformly converted into scene graphs. Thus, we propose SceneDiffusion, a scene generation model embedded with spatial constraints that parses object and relational information from scene graphs as conditional signals to guide the generation process, ultimately producing scene images that adhere to spatial distribution requirements.

The main contributions of our work are manifested in the following three aspects:

- A grid-based spatial scene representation framework that integrates the characterization of objects with topological, directional, and distance-based relationships into a unified formalized description. This framework bridges qualitative and quantitative spatial representations, adapting to diverse application needs.
- A deep neural network embedded with spatial constraints for scene generation. By introducing a Spatial relationship Attention-aware Graph (SAG) module, the model adaptively learns spatial layouts under topological, directional, and distance constraints, enhancing its spatial comprehension. Further, a Layout information-guided

Conditional Diffusion (LCD) module is incorporated to generate visually detailed and semantically consistent scene images.

- A purpose-built spatial scene dataset containing relational descriptions and geographic semantics. The dataset includes both structured semantic representations (e.g., object attributes, relational predicates) and corresponding rasterized visual representations, enabling rigorous evaluation of the proposed method.

This study employs generative AI to tackle scene generation under spatial constraints. On one hand, it offers a relational representation paradigm for geospatial cognitive modeling; on the other, it provides a generative framework for scene simulation. The work demonstrates practical potential in applications such as emergency response and virtual geographic environments, holding significant theoretical and applied value.

## 2. Related Work

### 2.1. Text-to-Image Generation

In recent years, text-to-image generation has emerged as a prominent research direction within the field of AI, garnering significant attention from scholars and research teams. This technology aims to automatically generate corresponding scene images by understanding and parsing textual descriptions. Current approaches in text-to-image generation can be categorized into three main technical routes:

**Scene Graph-Based Image Generation.** Scene graphs provide structured semantic representations of scenes. For instance, Johnson et al. proposed Sg2Im, an image generation model based on graph neural networks, which processes scene graph inputs to predict bounding boxes and segmentation masks for objects, subsequently generating scene layouts and refining them using a cascaded refinement network [7]. Ashual et al. further enhanced the diversity of generated images by decoupling object appearance embeddings from layout embeddings in scene graphs and introducing random vectors before generating object masks [8]. Vo et al. predicted visual relation layouts based on subject-predicate relationships in scene graphs and rendered the final scene images using these layouts [9]. Yang et al. introduced SGDiff, a model that employs masked autoencoder pretraining to learn local scene graph features and contrastive learning pretraining to capture global information, thereby improving alignment between scene graphs and images. They then constructed a latent diffusion model to generate images from scene graphs [10].

**Two-Stage Generation: Text-to-Layout-to-Image.** For example, Lian et al. proposed a scene generation method leveraging large language models (LLMs), where the LLM first generates a scene layout from the input text prompt, followed by image synthesis via Stable Diffusion [11] conditioned on the layout [12]. Wang et al. introduced a text-to-image model that predicts bounding boxes for entities mentioned in the text, using these spatial cues to guide the diffusion model's inference process, ultimately improving the quality of generated scenes [13].

**Direct Text-Encoded Image Generation.** For example, Feng et al. enhanced Stable Diffusion by integrating CLIP-based [14] text encoding and cross-attention mechanisms to facilitate interaction between image features and text embeddings, optimizing compositional semantics in scene generation [15]. Chefer et al. introduced Generative Semantic Nursing (GSN), which strengthens attention to all subject entity tokens in text prompts via an improved cross-attention mechanism, thereby enhancing semantic accuracy in generated scenes [16]. Yang et al. incorporated positional tokens alongside textual encodings to quantify spatial coordinates of objects, enabling region-controlled scene generation [17].

While these methods have achieved remarkable progress in image quality and content relevance, they exhibit limitations in handling geospatial scenes and deep spatial relationship understanding. Notably, although scene graph-based approaches effectively

utilize spatial relationships between objects, current research remains confined to simple geometric relations (e.g., “right of,” “below”) and employs relatively uniform processing strategies. More complex spatial relationships—such as distance, topology, and directional relations—have yet to be thoroughly explored, despite their critical role in accurately generating spatial scenes.

To address these gaps, this study systematically models complex spatial relationships among objects and introduces a SAG module to explicitly encode relational information, thereby enhancing the model’s ability to understand and control object spatial distributions.

## 2.2. Diffusion Models

Currently, diffusion models have gained increasing popularity due to their stable training processes and superior generative capabilities [18–21]. These models simulate physical diffusion mechanisms, employing a progressive strategy to transform random noise into a target data distribution, thereby achieving high-quality scene image synthesis. The generation process consists of two key phases: the forward diffusion process gradually corrupts real data into pure noise by adding Gaussian noise, and the reverse denoising process reconstructs the original image by iteratively removing noise using a neural network. In 2020, Ho et al. proposed Denoising Diffusion Probabilistic Models (DDPM) [22], marking the first successful application of this technique to image generation tasks and laying a critical foundation for subsequent research. Currently, diffusion models primarily employ two conditioning mechanisms for controlled generation: classifier guidance and classifier-free guidance.

**The classifier guidance mechanism** involves training an explicit classifier and incorporating it as a condition into diffusion models, utilizing backpropagation for classifier gradient descent [23]. Kawar et al. replaced traditional classifiers with a time-dependent adversarially robust classifier to guide the generation of semantically consistent scene images [24]. Shenoy et al. employed a pre-trained classifier to guide adaptive sampling at each timestep, improving the classification accuracy of generated images [25]. Although the classifier guidance mechanism is computationally efficient, its generative performance is constrained by the classifier’s capability.

**The classifier-free guidance mechanism**, on the other hand, directly encodes conditional information during the training process of diffusion models [26], enabling the models to learn richer semantic representations. Yang et al. introduced a spatial dependency parser to encode object-level spatial semantic consistency as layout embeddings, generating scenes with perceptually harmonious object styles and contextual relationships [27]. Baykal et al. integrated prototype learning into diffusion models by training a prototype codebook and combining temporal encodings with the pre-trained codebook as conditioning signals to guide the diffusion process [28].

As a powerful generative framework, diffusion models excel at modeling the transition from simple to complex distributions. Leveraging this technical advantage, we explore the application of spatial layout-guided classifier-free conditional diffusion models for scene generation. Our approach aims to enhance the model’s ability to synthesize spatially accurate and semantically consistent spatial scenes, addressing the unique challenges of scene generation in GeoAI.

## 3. Method

### 3.1. Preliminaries

**Diffusion Models.** Diffusion models are generative models grounded in probability theory, primarily divided into a forward diffusion process and a reverse denoising process. Specifically, the forward diffusion process involves gradually adding Gaussian

noise  $\epsilon \sim N(0, 1)$  to the original data  $x_0$  until it evolves into pure noise  $x_T$ . This process is typically represented by a Markov chain, implying that the data at a certain time  $t$  is solely dependent on the data from the previous time step  $t - 1$ , with a transition probability given by:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \forall t \in \{1, \dots, T\}, \quad (1)$$

where  $t$  denotes the time step,  $T$  is the total number of steps, and  $\beta_t$  controls the amount of noise added at the  $t$ -th time step.  $I$  represents the identity matrix with the same dimensions as the original data. Since the noise added progressively follows a Gaussian distribution, the forward diffusion process does not possess any trainable parameters. Through the technique of reparameterization, Equation (1) can be written as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I), \quad (2)$$

where  $\alpha_t = 1 - \beta_t$ , and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ .

The reverse denoising process takes Gaussian noise  $x_T$  as input and trains a denoising neural network to predict and eliminate the noise added at each step, ultimately restoring the original data. The reverse process share a similar iterative structure with the forward process, with the formula expressed as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (3)$$

where  $p_\theta$  denotes the trained denoising network, with  $\theta$  representing the parameters of the network. The mean  $\mu_\theta(x_t, t)$  and covariance  $\Sigma_\theta(x_t, t)$  are parameterized by learnable neural networks. In practice, the training objective of diffusion models is typically set as the mean square error (MSE) of noise prediction, formulated as follows:

$$\mathcal{L} = \mathbb{E}_{x_0, \epsilon, t} [||\epsilon - \epsilon_\theta(x_t, t)||^2], \quad (4)$$

where  $\epsilon$  is the noise added during the forward process, and  $\epsilon_\theta$  is the noise predicted by the neural network. By minimizing the noise prediction error, the model effectively learns the underlying data distribution characteristics, thereby achieving superior performance in generation tasks.

**Spatial Relations.** In the field of Geographic Information Science (GIS), spatial relations constitute a theoretical framework with a well-defined logical structure, encompassing three fundamental dimensions: topological relations, directional relations, and distance relations.

Topological relations delineate the invariant properties among spatial objects under transformations such as rotation or uniform scaling, reflecting the most fundamental structural associations between geographic entities. Among existing models of spatial topology, the Region Connection Calculus basic 8 (RCC8) proposed by Randell et al. has been particularly influential due to its completeness [29]. Based on the assumption of indivisible regions, the RCC8 model defines eight independent and exhaustive primitive relations: Disconnected, Externally Connected, Partially Overlapping, Equals, as well as directional inverse relation pairs—Tangential Proper Part and its inverse, and Non-Tangential Proper Part and its inverse.

Directional and distance relations typically involve both qualitative and quantitative representations. Directional relations characterize the orientation between spatial objects and depend on three key elements: the primary object, the reference object, and the reference coordinate system. Quantitative descriptions are achieved by measuring the angle between the line connecting the primary and reference objects and the axes of the reference system. For qualitative representation, widely accepted models include 4-directional (North, South,

West, East) or 8-directional (adding Northwest, Southwest, Northeast, Southeast) systems based on projection or conical partitioning.

Distance relations describe the spatial proximity between objects. Quantitative expressions employ precise metric values, while qualitative representations exhibit multi-granularity characteristics, with different classification standards applied depending on the specific use case.

In summary, spatial relationships, through the three interrelated yet independent components of topology, direction, and distance, collectively provide a comprehensive characterization of spatial structure. This integrated framework enables robust spatial reasoning and supports complex geospatial analyses.

### 3.2. Spatial Scene Representation Framework

Spatial scene representation is one of the core issues in GIS, aiming to provide formalized descriptions and representations of geographic objects and their interrelationships. Current approaches typically employ separate computational frameworks to handle object characterization along with topological, directional, and distance relationships, lacking unified definitions and computational models. Moreover, due to the complexity of spatial relationship calculations, efficient spatial information modeling and analysis remains challenging. To address these limitations, we propose a grid-based spatial scene representation framework that unifies the characterization of geographic objects with the computation of all three spatial relationships within an integrated paradigm. This framework transforms complex spatial relationship computations into simplified operations on spatial grid indices and sets, achieving computationally efficient holistic representation.

Grid is a regular spatial partitioning method that divides continuous geographic space into uniform cells, each uniquely identified and rapidly processed through grid indexing. This framework integrates qualitative and quantitative spatial representations. On one hand, it discretizes spatial scenes by converting continuous object geometries and their distributions into grid cells, providing a generalized approach for object and relationship description. On the other hand, grid cell encoding precisely records the coordinate ranges of objects, supporting quantitative spatial relationship computation. The grid size determines the scale of spatial representation: smaller grids yield higher precision (quantitative dominance), while larger grids introduce ambiguity (qualitative dominance). By adjusting grid resolution, the framework dynamically balances qualitative and quantitative representations to accommodate diverse application scenarios.

In GIS, spatial objects are typically abstracted as points, lines, or polygons, whereas our framework unifies all geographic entities as collections of occupied grid cells to achieve consistent representation. For operational simplicity, we encode the 2D grid array using row ( $G_y$ ) and column ( $G_x$ ) indices, where point objects are represented by a single grid cell, and line/polygon objects are characterized by multiple connected grid cells. The size of an object is approximately represented by the number of grids it occupies. This unified representation enhances computational efficiency while maintaining expressiveness.

The topological, directional, and distance relationships between objects are derived through grid cell indexing and set operations. Consider two objects  $A$  and  $B$  in a scene, where object  $A$  consists of  $m$  grid cells denoted as  $Obj_A = \{G_1^A, G_2^A, \dots, G_m^A\}$ , and object  $B$  consists of  $n$  grid cells denoted as  $Obj_B = \{G_1^B, G_2^B, \dots, G_n^B\}$ , with  $G$  representing a single grid cell. Based on the sets  $Obj_A$  and  $Obj_B$  and their constituent grid cell indices, the three types of spatial relationships between objects  $A$  and  $B$  are extracted as follows:

Topological relationships are determined through containment analysis between grid cell sets. Drawing upon the classical RCC model, we define five fundamental topological relationships for spatial scenes: Disjoint, Overlap, Contains, Inside, and Equal, as illustrated

in Figure 1. The specific extraction methodology for topological relationships is detailed in Appendix A.

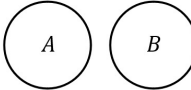
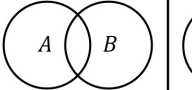
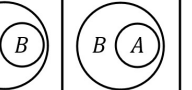
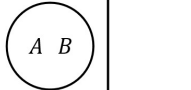
| Topological Relations | Disjoint                                                                          | Overlap                                                                            | Contains                                                                            | Inside                                                                              | Equal                                                                               |
|-----------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Illustration          |  |  |  |  |  |

**Figure 1.** The five topological relationships of object *A* relative to object *B*.

Directional relationships are extracted based on the coordinate positions (i.e., index values  $G_x$  and  $G_y$ ) of all grid cells comprising the objects. The directional relationship of the primary object relative to the reference object is determined by averaging the directional relationships between each grid cell of the primary object and each grid cell of the reference object. Specifically, the directional value  $Dir(A, B)$  of spatial object *B* relative to spatial object *A* is calculated through the following formula:

$$Dir(A, B) = \frac{1}{m} \sum_{i=1}^m \left( \frac{1}{n} \sum_{j=1}^n Dir(G_i^A, G_j^B) \right), \quad (5)$$

where  $Dir(G_i^A, G_j^B)$  represents the directional value of grid cell  $G_j^B$  of object *B* relative to grid cell  $G_i^A$  of object *A*, and  $m$  and  $n$  respectively represent the number of grids occupied by objects *A* and *B*.

To align with human spatial cognition, we employ an 8-direction system (including North-*N*, Northeast-*NE*, East-*E*, Southeast-*SE*, South-*S*, Southwest-*SW*, West-*W*, Northwest-*NW*) augmented with a central direction (denoted as *C*) to characterize directional relationships between grid cells. This forms a  $3 \times 3$  directional matrix where each element's value ranges between  $[0, 1]$ , with the sum of all elements equaling 1. The calculation formula for directional matrix  $Dir(G_i^A, G_j^B)$  is as follows:

$$Dir(G_i^A, G_j^B) = \begin{bmatrix} \mu_{NW}(G_j^B) & \mu_N(G_j^B) & \mu_{NE}(G_j^B) \\ \mu_W(G_j^B) & \mu_C(G_j^B) & \mu_E(G_j^B) \\ \mu_{SW}(G_j^B) & \mu_S(G_j^B) & \mu_{SE}(G_j^B) \end{bmatrix}, \quad (6)$$

where each matrix element  $\mu(G_j^B)$  indicates the membership degree of grid cell  $G_j^B$  within various qualitative directional regions relative to reference grid cell  $G_i^A$ , with specific computational methods provided in Appendix A.

Distance relationships are computed using either Euclidean or Manhattan distance between grid cells, with the grid's regularity enabling efficient distance calculations. Specifically, the distance between objects *A* and *B* can be defined as the minimum distance between their constituent grid cells:

$$Dist(A, B) = \min_{G_i^A \in Obj_A, G_j^B \in Obj_B} d(G_i^A, G_j^B), \quad (7)$$

where  $d(G_i^A, G_j^B)$  represents the distance between grid cells  $G_i^A$  and  $G_j^B$ . We adopt Manhattan distance as the metric, calculated as shown in the following formula:

$$d(G_i^A, G_j^B) = |G_{j,x}^B - G_{i,x}^A| + |G_{j,y}^B - G_{i,y}^A|. \quad (8)$$

This metric not only features low computational complexity but also adapts well to regular grid structures, demonstrating high practicality in spatial scene representation.

In summary, the grid-based spatial scene representation framework integrates both the characterization of objects and the extraction of relationships into a unified model, providing an efficient and flexible solution for spatial information modeling and analysis.

### 3.3. SceneDiffusion

The scene generation model proposed in this study adopts a two-stage architecture, as illustrated in Figure 2. In the first stage, the model takes a scene graph containing objects along with their topological, directional, and distance relationships as input. Through embedding layers, it encodes both object features and relationship information, while utilizing the SAG module to adaptively learn the influence mechanisms of three spatial relationships on object layout, thereby generating the scene layout constrained by spatial relationships. In the second stage, the model employs the LCD module to progressively denoise and generate scenes conforming to the target distribution, based on the layout output from the first stage. Through the coordinated operation of these modules, SceneDiffusion effectively integrates spatial constraints to generate scene images that maintain semantic consistency with the input scene graph.

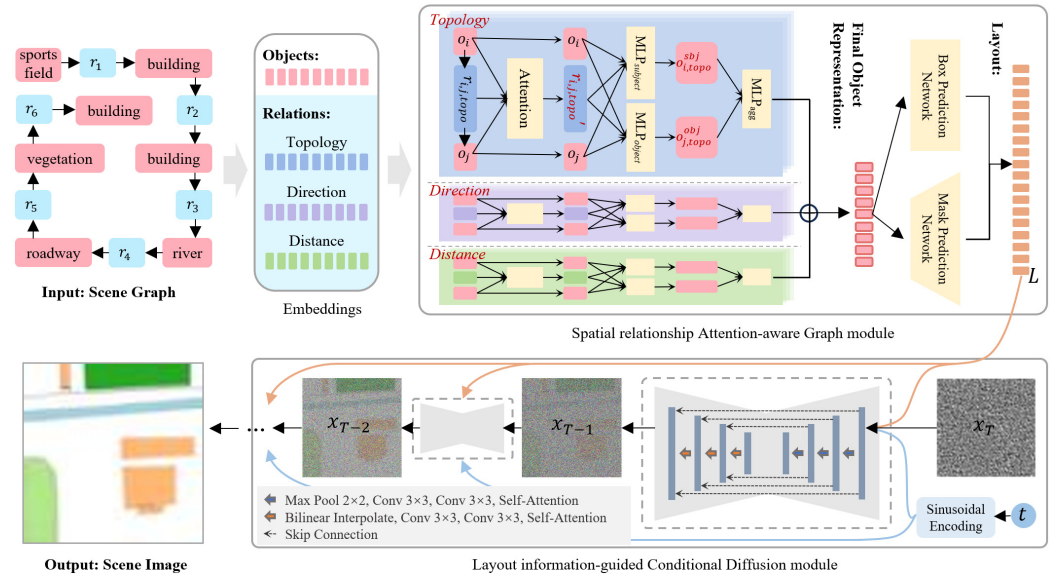


**Figure 2.** The architecture of SceneDiffusion. In the output scene image, light green represents the sports field, dark green represents vegetation, blue represents rivers, orange represents buildings, and black lines represent roadways.

**Scene Graph.** The model input employs a structured scene graph representation, defined as a collection of spatial triplets  $SG = (Obj_i, R_{i,j}, Obj_j)$ . Here,  $Obj_i, Obj_j \in O^C$  represent objects with category information, where  $C$  denotes the total number of object categories;  $R_{i,j} = \langle Topo(i, j), Dir(i, j), Dist(i, j) \rangle$  represents the topological, directional, and distance relationships of object  $Obj_j$  relative to object  $Obj_i$ .  $Topo(i, j)$  uses one-hot encoding to represent five fundamental topological relationships,  $Dir(i, j)$  is a  $3 \times 3$  directional matrix describing relative orientation between objects, and  $Dist(i, j)$  represents one or more distance metrics between objects. This triplet representation comprehensively captures spatial associations among objects in the scene, providing structured input for subsequent spatial layout and scene generation, thereby supporting precise spatial relationship modeling.

**Embeddings.** During the model's preprocessing stage, a set of learnable encoding layers maps discrete symbolic representations in the scene graph to a unified

dense vector space. Specifically, for any object  $Obj_i$ , its category information and size-indicating grid count are embedded separately, then concatenated to form the feature vector  $o_i \in \mathbb{R}^{d_m}$ , where  $d_m$  denotes the feature dimension. For spatial relationships  $Topo(i, j)$ ,  $Dir(i, j)$ , and  $Dist(i, j)$ , separate embedding layers transform them into vectors  $r_{i,j,topo}, r_{i,j,dir}, r_{i,j,dist} \in \mathbb{R}^{d_m}$  respectively. This embedding approach not only preserves the structural information of the scene graph but also converts it into vector representations suitable for neural network processing, establishing the foundation for feature learning in subsequent modules. Moreover, the parameters of embedding layers are optimized through end-to-end training, enabling adaptive capture of semantic features for both objects and relationships.

### 3.3.1. Spatial Relationship Attention-Aware Graph Module

To comprehensively model spatial distribution patterns within scenes and effectively propagate dependencies among objects associated through multiple spatial relationships, we design the SAG module. This module employs three parallel processing pathways to respectively model object representations constrained by topological, directional, and distance relationships. The design inspiration stems from the concept of aggregating node neighborhood information through edge features in graph convolutional networks. However, unlike conventional graph convolution approaches, spatial relationships as multidimensional complex edge features exhibit distinct influence mechanisms of their topological, directional, and distance attributes on object distribution in scenes. This heterogeneity precludes the use of a unified processing function.

Based on these observations, we introduce attention mechanisms to adaptively learn objects' varying attention levels to different spatial relationships, thereby achieving information propagation and updating under multi-dimensional spatial constraints. Specifically, at each layer within the SAG module, for a given spatial triplet (*subject, relation, object*), the relationship representations are first updated through attention mechanisms to capture influence weights of different relationships on objects distribution. Subsequently, information propagation for both the *subject* and the *object* is computed under each relationship constraint. Finally, by fusing information from all relationship pathways, updated vector representations of objects are generated. This design not only effectively distinguishes the roles of different spatial relationships but also enhances the model's capability to represent complex spatial distribution patterns, thereby providing more precise spatial constraints for subsequent scene layout generation.

Taking topological relationship constraints as an example, for the vector representation  $(o_i, r_{i,j,topo}, o_j)$  of a spatial triplet, the *subject* vector  $o_i$  is mapped as Query, the *object* vector  $o_j$  as Key, and the topological relationship vector  $r_{i,j,topo}$  as Value. The information updating of topological relationships is implemented through an attention mechanism with the following computational process:

$$r'_{i,j,topo} = \text{softmax}\left(\frac{o_i W_q (o_j W_k)^T}{\sqrt{d_m}}\right) \cdot r_{i,j,topo}, \quad (9)$$

where  $W_q$  and  $W_k$  are learnable parameter matrices. With this design, the model can dynamically adjust the representation of topological relationships based on the specific semantics of objects in spatial triplets, thus more accurately capturing the association characteristics of different objects in the topological space.

The vector updating process of objects is relatively complex, primarily due to two factors: firstly, a single object may simultaneously participate in multiple spatial relationships; secondly, the same object may serve as either the *subject* or the *object* in different spatial triplets, with significant semantic role differences. To effectively handle this complexity,

our model employs a Multi-Layer Perceptron (MLP) for stepwise computation of object update vectors. Specifically, under topological constraints, the vector representation of object  $Obj_i$  as the *subject* is:

$$o_{i,topo}^{sbj} = \{\text{MLP}_{subject}(\text{concat}(o_i, r'_{i,j,topo}, o_j))\}, \text{ if } (Obj_i, \text{Topo}(i, j), Obj_j) \in SG, \quad (10)$$

and as the *object* is denoted as:

$$o_{i,topo}^{obj} = \{\text{MLP}_{object}(\text{concat}(o_j, r'_{i,j,topo}, o_i))\}, \text{ if } (Obj_j, \text{Topo}(j, i), Obj_i) \in SG. \quad (11)$$

Through average pooling and MLP-based aggregation of information when object  $Obj_i$  serves as both the *subject* and the *object*, its updated vector representation under topological constraints is obtained:

$$o'_{i,topo} = \text{MLP}_{agg}(\text{average}(o_{i,topo}^{sbj}, o_{i,topo}^{obj})). \quad (12)$$

Subsequently, by stacking multiple neural network layers and repeating the aforementioned updating processes for both topological relationship vectors and spatially constrained object vectors, topological information propagates across multi-level neighborhoods. This hierarchical propagation mechanism effectively captures long-range dependency relationships among objects in scenes.

Similarly, for directional and distance relationship constraints, the model employs identical multi-layer neural networks and attention mechanisms to compute updated representations of the directional relationship vector  $r_{i,j,dir}$  and distance relationship vector  $r_{i,j,dist}$ , respectively. This process yields the updated vector representation  $o'_{i,dir}$  of object  $Obj_i$  under directional constraints and  $o'_{i,dist}$  under distance constraints. Through this parallel processing mechanism, the model achieves effective information propagation and fusion under different spatial relationship constraints. Ultimately, the final vector representation of object  $Obj_i$  is obtained through summation operation:

$$o'_i = o'_{i,topo} + o'_{i,dir} + o'_{i,dist}. \quad (13)$$

Based on the final vector representations of objects, the model further computes scene layout information to conditionally generate the final image. The scene layout consists of layout information from all objects in the scene, with each object's layout comprising both bounding box and mask components. Specifically, the bounding box information  $b_i \in \mathbb{R}^4$  for object  $Obj_i$ , encompassing coordinates of two diagonal vertices, is generated by a MLP-based bounding box prediction network  $B$ :

$$b_i = B(o'_i). \quad (14)$$

The bounding box prediction network employs multiple fully-connected layers with nonlinear activation functions to map object vector representations into four-dimensional bounding box coordinates, thereby precisely describing each object's spatial position and extent within the scene. The mask information  $m_i \in \mathbb{R}^{N_m \times N_m}$  for each object is generated by a mask prediction network  $M$ :

$$m_i = M(o'_i). \quad (15)$$

The mask prediction network progressively restores spatial resolution through up-sampling operations while incorporating convolutional layers to extract local features, ultimately producing binary masks that match object shapes. To align mask information with bounding boxes, the model performs element-wise multiplication between the object vector

$o'_i$  and predicted mask  $m_i$ , followed by bilinear interpolation to map it onto the spatial extent of predicted bounding box  $b_i$ . This yields the layout information  $L_i \in \mathbb{R}^{d_m \times h \times w}$  for object  $Obj_i$ , where  $h$  and  $w$  represent the height and width of the predicted image, respectively. Finally, the model aggregates layout information from all objects in the scene to produce the scene layout representation  $L$ , which serves as conditional input for subsequent generation modules.  $L$  not only encapsulates spatial position and shape information for all objects in the scene but also preserves semantic features through object vector representations  $o'_i$ , thereby providing both spatial and semantic constraints for scene generation.

### 3.3.2. Layout Information-Guided Conditional Diffusion Module

The classifier-free guided conditional diffusion model offers an efficient and flexible framework for incorporating conditional information into diffusion models. Building upon this foundation, we construct the LCD module, which is designed to utilize the scene layout information  $L$  generated by the SAG module as conditional input to control the scene generation process, thereby guiding the diffusion model to produce scenes that conform to spatial distribution constraints. Specifically, considering that the noisy image  $x_T$  introduced during the generation process of the diffusion model shares the same spatial dimensionality as the target image, we directly concatenate the scene layout information  $L$  with the noisy image  $x_T$  along the channel dimension and input it into the noise estimator. Traditional classifier-free guidance methods necessitate the simultaneous training of both conditional and unconditional models, resulting in a large number of parameters and higher computational complexity. To simplify implementation and enhance computational efficiency, we adopt a lightweight strategy: treating the unconditional generation process as a special case of conditional generation when the conditional information is null. This process can be formalized as:

$$\tilde{\epsilon}_\theta(x_t, t|L) = \lambda \epsilon_\theta(\text{concat}(x_t, L), t) + (1 - \lambda) \epsilon_\theta(x_t, t|\emptyset), \quad (16)$$

where,  $\lambda \geq 1$  serves as a hyperparameter controlling the strength of the conditional information;  $\epsilon_\theta$  is a neural network based on the U-Net architecture, responsible for estimating the noise;  $x_t$  represents the noisy image at time step  $t$ ;  $L$  denotes the scene layout information; and  $\emptyset$  indicates the unconditional diffusion process. The model is trained with the objective of minimizing noise prediction error, as expressed in Equation (4).

The layout information  $L$  contains spatial attributes, bounding boxes, and mask information for all objects in the scene. Through the conditional diffusion model, this information is progressively decoded into image content that adheres to actual spatial distributions, ensuring the final generated scenes satisfy spatial relationship constraints among objects. The advantage of this conditional guidance mechanism lies in its dual capability: it preserves the diffusion model's inherent ability to generate high-quality images while simultaneously controlling the spatial distribution of objects through layout information.

Furthermore, the hyperparameter  $\lambda$  offers flexibly control over the degree of dependence of the generated images on layout conditions. When  $\lambda$  is relatively large, the generated images strictly adhere to the constraints imposed by the layout information  $L$ ; whereas, when  $\lambda$  is smaller, the model relies more heavily on the unconditional diffusion process, resulting in the generation of more diverse images. This flexibility enables the model to adapt to various scenario requirements, further enhancing its practicality and robustness in real-world applications.

## 4. Experiment

### 4.1. Geospatial Scene (GS) Dataset

To effectively validate the scene generation model, we constructed a high-fidelity training dataset based on GIS data models and human spatial cognition principles. This dataset, extracted from vector map data, contains paired samples of scene graphs and corresponding images. The vector map data is based on the WGS 1984 Web Mercator projection coordinate system, which includes three types of geographic elements: points, lines, and polygons, providing comprehensive spatial information and covering rich geographic spatial features. The design of the dataset follows three core principles: (1) Geographic accuracy—preserving geometric and attribute characteristics of vector maps to ensure precise spatial representation; (2) Semantic completeness—covering three types of geographic features (points, lines, and polygons) along with their topological, directional, and distance relationships; (3) Cognitive plausibility—scene graph structures reflecting spatial semantics that align with human descriptive patterns of spatial environments. The dataset construction process involves three key steps: systematic selection of scene samples, scene graph generation using our spatial representation framework and graph algorithms, and corresponding image generation through GIS processing tools. By establishing a mapping between scene graphs and images, this dataset provides multimodal training data with comprehensive spatial constraints.

**Spatial scene sampling.** We employed a 500 m × 500 m sliding window to clip original vector map data. Drawing from existing scene generation research [7], we selected map segments as study samples. These segments contain 3–8 polygonal objects with at least two distinct types, ensuring each scene maintains appropriate complexity and diversity.

**Scene graph generation.** This process transforms geographic elements into gridded representations using our spatial representation framework. To simulate human cognitive and descriptive patterns of spatial environments, which tend to focus on adjacent objects rather than exhaustive combinations, we constructed graph structures with objects as nodes and relationships as edges. Using Prim’s minimum spanning tree algorithm [30] with minimum grid-based distances as edge weights, we generated connected graphs with minimal total weights as structured scene graph representations, balancing object accessibility with human spatial cognition principles.

**Scene images generation.** This process is implemented using QGIS at 1:5000 scale, applying customized styles for different objects (e.g., buildings, rivers, vegetation) to ensure accurate spatial representation and visual clarity. PNG-format scene images were exported based on predefined spatial boundaries.

Figure 3 demonstrates a sample pair of “scene graph-image”. In the scene graph on the left, red rectangles represent objects and blue rectangles denote topological, directional and distance relationships. For clarity, directional relationships are simplified to dominant directions (maximum values in  $3 \times 3$  directional matrices), while distance relationships are quantified as Manhattan distances between object centroids. The marked spatial triplet in the scene graph indicates vegetation located south of a sports field, with disconnected topology and 183-unit centroid distance, which is consistent with the actual spatial distribution shown in the corresponding scene image on the right.

Using vector map data from a certain city, we ultimately constructed a high-quality geospatial scene dataset containing 5964 sample pairs through this processing pipeline. We divided the dataset into training, testing, and validation subsets in an 8:1:1 ratio to facilitate robust model training and evaluation. Unlike existing datasets for image generation that contain simple geometric relationships in perspective-view photographs, our dataset specifically focuses on geospatial relationships in map-view scenes. It uniquely integrates topological, directional, and distance relationships, providing reliable data support for research on scene generation embedded with spatial constraints.

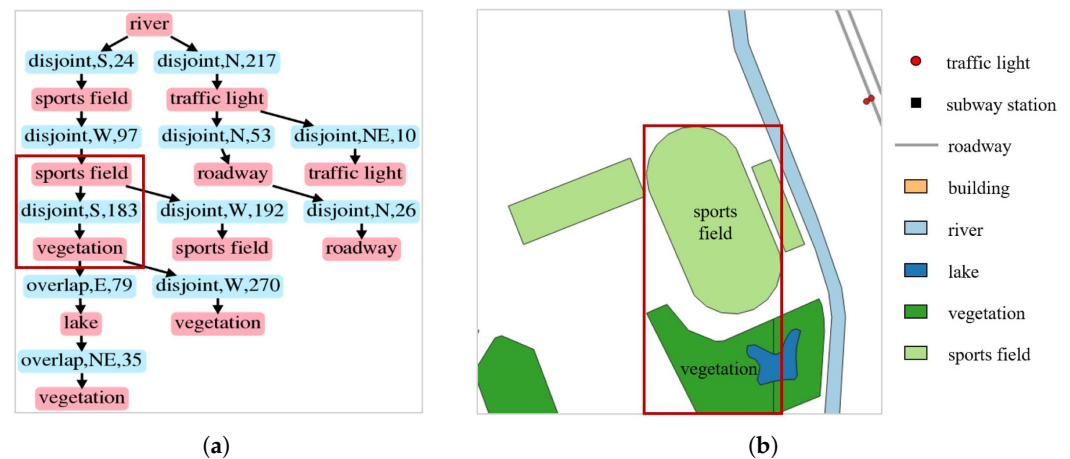


**Figure 3.** A “scene graph-image” sample pair. (a) Scene graph, (b) Scene image.

#### 4.2. Evaluation Metrics

To comprehensively evaluate the performance of the scene generation model, we employ three complementary evaluation metrics that assess different aspects of generation quality. These metrics are selected based on their established validity in computer vision and specific relevance to spatial scene evaluation.

Fréchet Inception Distance (FID) [31] is a widely adopted metric in the field of image generation, which assesses the quality of generated images by comparing the distribution differences between generated and real images in feature space. A lower FID value indicates superior generation quality.

Structural Similarity Index Measure (SSIM) [32] evaluates the similarity between generated and real images from three dimensions: luminance, contrast, and structure. A higher SSIM value denotes greater similarity, effectively reflecting the model’s capability to preserve the structural integrity of spatial scenes.

Peak Signal-to-Noise Ratio (PSNR) [32] quantifies the detail fidelity of generated images from the perspective of reconstruction accuracy by computing pixel-level errors, with higher values indicating superior image fidelity.

These three metrics collectively provide a multidimensional assessment of the generative model’s performance at the feature, structural, and pixel level, thereby offering a comprehensive validation of the model’s proficiency in spatial information understanding and the quality of scene generation.

#### 4.3. Experimental Details

The experiments were conducted on an NVIDIA GeForce RTX 3090 GPU, with the computational environment implemented using CUDA 11.3, Python 3.9, and PyTorch 1.12 to ensure compatibility and computational efficiency.

The key parameters for model training are configured as follows: The initial learning rate is set to  $3 \times 10^{-4}$ , with a batch size of 12, generating scene images of size  $64 \times 64$  pixels. The diffusion process employs a total of  $T = 1000$  steps, with the noise schedule  $\beta_t$  linearly increasing from  $1 \times 10^{-4}$  to 0.02. The model dimension  $d_m$  is set to 128, and the training is conducted over 500 epochs. This parameter configuration ensures adequate model learning while promoting convergence. For baseline models, all parameters are set according to the recommended values in their original publications.

#### 4.4. Qualitative and Quantitative Results

We conduct comparative experiments against two representative baselines: Sg2Im [7] and SGDiff [10]. Sg2Im represents a classical approach for scene-graph-to-image generation,

which utilizes an end-to-end training framework with explicit scene graph encoding to preserve semantic consistency between generated images and input scene graphs. In contrast, SGDiff is a recently proposed diffusion-based image generation method that achieves high-quality image synthesis from scene graphs through a two-stage training paradigm. By comparing these two baseline models with distinct methodological paradigms, we comprehensively evaluate the relative performance of our proposed SceneDiffusion across different technical approaches.

Our evaluation on the GS dataset incorporates both qualitative and quantitative assessments to validate the model's effectiveness in scene generation tasks. For fair experimental comparison, baseline models utilize the joint embeddings of topological, directional, and distance relationships from the GS dataset as relational inputs.

#### 4.4.1. Qualitative Result Analysis

Figure 4 presents generation results from different models across diverse spatial scenes. The object categories represented by different colors are shown in Figure 3. These scenes are carefully selected to include both simple and complex spatial relationship configurations, incorporating various object types such as roads, vegetation, and buildings. Each scene graph explicitly specifies the topological, directional, and distance relationships among objects within the scenes.

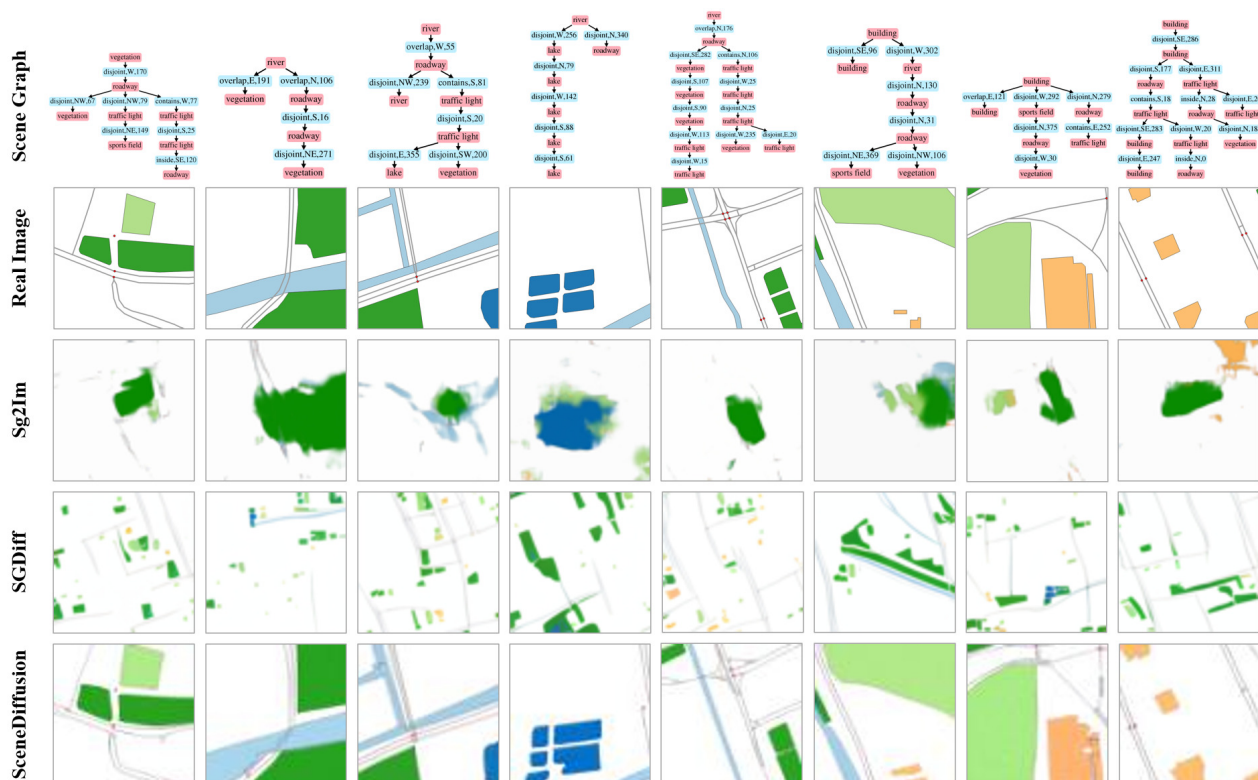


**Figure 4.** Comparison of qualitative results. The first row displays the input scene graph, while the second row presents the corresponding ground truth image. The subsequent three rows sequentially showcase the generated results from Sg2Im, SGDiff, and SceneDiffusion respectively.

Experimental results demonstrate that our SceneDiffusion significantly outperforms the baseline methods in terms of both visual quality and semantic consistency. Specifically, our model faithfully reconstructs objects while maintaining their boundary clarity and shape integrity. More importantly, the spatial relationships between objects (e.g., relative positions of buildings) in the generated scenes consistently align with the descriptions in the input scene graphs, demonstrating the model's superior capability in spatial relationship modeling.

In comparison, the baseline models exhibit limitations. Sg2Im produces blurred outputs with poor detail representation, showing low semantic consistency with the input scene graphs. Its generated results frequently contain positional errors among objects and fail to properly reflect the spatial relationship constraints specified in the scene graphs. While SGDiff generates more realistic textures, it still fails to correctly learn the objects and their relationships. The generated scenes often contain incorrect numbers of objects that mismatch the scene graph descriptions, and the spatial relationships between objects are frequently misrepresented.

Furthermore, we validate the model's responsiveness by introducing incremental modifications to the scene graph. Specifically, in Figure 5, we progressively reduce the spatial distance between the sports field object and the lake object in the input scene graph, while maintaining all other spatial relationships. The generated scene images (left to right) demonstrate corresponding gradual decreases in inter-object distance, confirming that the output images accurately reflect the adjusted spatial constraints in the scene graph.

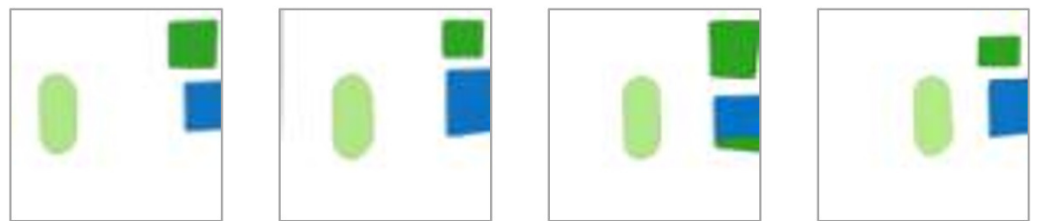


**Figure 5.** Experimental results on SceneDiffusion for changing the distance relationship in a scene graph.

These results confirm that SceneDiffusion effectively utilizes the spatial relationship information in scene graphs during the generation process, thereby validating the effectiveness of our model design.

#### 4.4.2. Quantitative Result Analysis

Table 1 presents the quantitative evaluation results of FID, SSIM, and PSNR metrics for different models on the GS dataset. The experimental results demonstrate that our proposed scene generation model significantly outperforms baseline models across all evaluation metrics. In terms of scene generation quality, SceneDiffusion achieves an FID score of 67.15, representing reductions of 56.6% and 64.6% compared to the Sg2Im (154.59) and SGDiff (189.47), respectively. This indicates that the scenes generated by our model are substantially closer to real data in terms of feature distributions, confirming its superior generation capability. Structural similarity assessment reveals that SceneDiffusion attains an SSIM value of 0.69, showing improvements of 35.3% and 60.5% over the Sg2Im (0.51) and SGDiff (0.43), respectively. These results demonstrate our model's remarkable advantage in preserving luminance, contrast, and structural characteristics of spatial scenes, as well as its ability to more accurately maintain the spatial layouts and inter-object relationships defined in the input scene graphs. Further validation of our model's superiority comes from PSNR comparisons. SceneDiffusion achieves a PSNR value of 17.80 dB, showing substantial improvement over both baseline models (Sg2Im: 11.12 dB, SGDiff: 11.95 dB), which indicates its enhanced performance in pixel-level reconstruction accuracy.

The comparison across all three metrics consistently demonstrates that SceneDiffusion exhibits clear advantages in generation quality, structural preservation, and reconstruction accuracy. The particularly notable improvement in FID metric strongly validates the model's capability to generate scene images whose feature distributions closely resemble real scenes. These quantitative findings align with the qualitative analysis conclusions, collectively proving the effectiveness of SceneDiffusion for scene generation tasks.

**Table 1.** Comparison of quantitative results.

| Models         | FID    | SSIM | PSNR     |
|----------------|--------|------|----------|
| Sg2Im          | 154.59 | 0.51 | 11.12 dB |
| SGDiff         | 189.47 | 0.43 | 11.95 dB |
| SceneDiffusion | 67.15  | 0.69 | 17.80 dB |

#### 4.5. Ablation Studies

To validate the contributions of core components in SceneDiffusion, we design four variant models for ablation studies. Table 2 presents the quantitative evaluation results of these ablation experiments. The *only Topo* variant retains solely topological relationship modeling while removing the processing pathways for directional and distance relationships; the *only Dir* variant preserves only directional relationship modeling; the *only Dist* variant maintains exclusively distance relationship modeling. Additionally, the *w/o Attn* variant retains all three relationship processing pathways but eliminates the attention mechanism in the SAG module, instead directly employing initial relationship vectors for object feature updates. This variant serves to verify the role of attention mechanisms in relational information fusion.

**Table 2.** Comparison of experimental results of SceneDiffusion variants.

| Variants         | FID    | SSIM | PSNR     |
|------------------|--------|------|----------|
| <i>only Topo</i> | 114.53 | 0.60 | 14.44 dB |
| <i>only Dir</i>  | 124.55 | 0.61 | 15.13 dB |
| <i>only Dist</i> | 126.72 | 0.53 | 10.40 dB |
| <i>w/o Attn</i>  | 103.37 | 0.64 | 15.58 dB |
| SceneDiffusion   | 67.15  | 0.69 | 17.80 dB |

The results reveal several key findings. The complete SceneDiffusion demonstrates superior performance across all metrics, particularly achieving a 35.0% FID improvement over the best-performing variant *w/o Attn*, confirming the synergistic effectiveness of multi-relational joint modeling coupled with attention mechanisms. Among single-relationship variants, the *only Topo* variant achieves the lowest FID value, indicating the fundamental importance of topological relationships in scene generation. The *only Dir* variant attains the highest SSIM score, highlighting directional relationships' particular significance in preserving structural layouts. The *only Dist* variant performs worst across all three metrics, suggesting distance constraints alone are insufficient for coherent scene generation. Beyond the aforementioned FID comparison, the *w/o Attn* variant shows a 7.2% reduction in SSIM and a 12.5% decrease in PSNR compared to the complete model, proving that the attention mechanism in the SAG module effectively captures dynamic influence weights across different relationships and plays a pivotal role in multi-relational information fusion.

These findings collectively validate our architectural design: (1) all three spatial relationship types contribute complementary benefits, and their combined modeling yields synergistic improvements beyond any single relationship, (2) the attention mechanism is essential for effectively integrating multi-relational constraints. The ablation results provide strong evidence for each component's necessity in achieving optimal scene generation performance.

## 5. Conclusions

This study presents SceneDiffusion, an innovative model for generating scenes with spatial constraints. The principal achievements can be summarized as follows: First, the grid-based representation framework successfully unifies qualitative and quantitative spatial representations, enabling extraction of topological, directional, and distance

relationships through adjustable grid resolutions. Second, the SAG module enhances relationship representations in scene graphs through attention mechanisms and updates object representations with spatial constraints, demonstrating exceptional capability in learning complex spatial dependencies. The LCD module effectively bridges consistent spatial constraints with high-quality visual generation, achieving state-of-the-art experimental results. Furthermore, to support the evaluation of our framework, we developed a geospatial scene dataset comprising scene graph-image pairs derived from vector map data, with spatial relationship descriptions and geographic semantics attached. It serves as a critical validation tool for our proposed approach and demonstrates the feasibility of integrating structured spatial knowledge into generative models.

The study achieves an organic integration of spatial information constraints and generative AI models. The attention-based multi-relational fusion mechanism significantly enhances the model's understanding of spatial scenes, providing a generalizable paradigm for spatial-aware generation tasks. In practical, the research outcomes can offer technical support for tasks such as emergency management, demonstrating broad application prospects. Regarding resolution, the clarity of generated scene images can be improved by enhancing the generative model itself. For instance, a cascaded diffusion framework could be adopted, where a low-resolution base model (the current SceneDiffusion) first synthesizes a globally coherent layout under spatial constraints, followed by a high-resolution refinement model to recover fine-grained details. Notably, higher-resolution generation inevitably demands greater computational power, necessitating a trade-off based on practical requirements.

Future work will focus on two primary directions: At the data level, we will incorporate multi-source heterogeneous data such as remote sensing imagery and expand geographical coverage to construct multimodal, cross-scale geospatial scene datasets. At the methodological level, we will explore more flexible and efficient spatial relationship representation approaches and investigate multimodal fusion mechanisms to achieve synergistic utilization of diverse spatial information including text, images, and vector data. These advancements will further strengthen the model's deployment potential in mission-critical applications. Furthermore, we plan to collaborate with domain experts in future work to validate the model's performance in practical applications.

**Author Contributions:** Conceptualization, Danhuai Guo and Shanshan Yu; methodology, Shanshan Yu, Jiaxin Zhu and Danhuai Guo; software, Shanshan Yu and Jiaqi Li; validation, Kai Wang and Jian Tu; formal analysis, Shanshan Yu, Jiaxin Zhu and Xunchun Li; investigation, Kai Wang and Jian Tu; resources, Danhuai Guo; data curation, Shanshan Yu and Jiaxin Zhu; writing—original draft preparation, Shanshan Yu; writing—review and editing, Danhuai Guo and Shanshan Yu; visualization, Jiaqi Li and Xunchun Li; supervision, Danhuai Guo; funding acquisition, Danhuai Guo. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China grant number 42371476, and the Fundamental Research Funds for the Central Universities of Beijing University of Chemical Technology grant number buctrc202132.

**Data Availability Statement:** The geospatial dataset comprises original vector maps collected for this study, which are not publicly available due to authorization restrictions from surveying and mapping authorities. Researchers can request access to the data by submitting a formal application to the corresponding author via email. Requestors will be required to agree to use the data solely for academic research purposes. The data will remain available for the foreseeable future upon reasonable request.

**Conflicts of Interest:** Author Jiaxin Zhu was employed by Beijing Sunwise Space Technology Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Appendix A

### Appendix A.1

**Table A1.** Extraction method for topological relationships.

| Topological Relationships | Descriptions                                                                                                                                                                                                     |
|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Disjoint                  | Object $A$ and object $B$ share no common grid cells, i.e., $Obj_A \cap Obj_B = \emptyset$                                                                                                                       |
| Overlap                   | Objects $A$ and object $B$ share at least one common grid cell while neither completely contains the other, i.e., $Obj_A \cap Obj_B \neq \emptyset$ , and $Obj_A \not\subseteq Obj_B, Obj_B \not\subseteq Obj_A$ |
| Contains                  | All grid cells of object $B$ form a proper subset of object $A$ 's grid cells, i.e., $Obj_B \subsetneq Obj_A$                                                                                                    |
| Inside                    | All grid cells of object $A$ form a proper subset of object $B$ 's grid cells, i.e., $Obj_A \subsetneq Obj_B$                                                                                                    |
| Equal                     | Objects $A$ and object $B$ occupy exactly the same grid cells, i.e., $Obj_A = Obj_B$                                                                                                                             |

### Appendix A.2

**Table A2.** Calculation method for directional relation membership degrees.

| Coordinate Positional Relationship between $G_j^B$ and $G_i^A$ | Value Range of $\tau$ | Membership Degree Calculation Formula                                 |
|----------------------------------------------------------------|-----------------------|-----------------------------------------------------------------------|
| $G_{j,x}^B > G_{j,x}^A, G_{j,y}^B > G_{j,y}^A$                 | $0 < \tau \leq 1$     | $\mu_{NE}(G_j^B) = \tau, \mu_E(G_j^B) = 1 - \tau$                     |
| $G_{j,x}^B > G_{j,x}^A, G_{j,y}^B < G_{j,y}^A$                 | $\tau > 1$            | $\mu_{NE}(G_j^B) = \frac{1}{\tau}, \mu_N(G_j^B) = 1 - \frac{1}{\tau}$ |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B < G_{j,y}^A$                 | $0 < \tau \leq 1$     | $\mu_{SE}(G_j^B) = \tau, \mu_E(G_j^B) = 1 - \tau$                     |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B > G_{j,y}^A$                 | $\tau > 1$            | $\mu_{SE}(G_j^B) = \frac{1}{\tau}, \mu_S(G_j^B) = 1 - \frac{1}{\tau}$ |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B < G_{j,y}^A$                 | $0 < \tau \leq 1$     | $\mu_{SW}(G_j^B) = \tau, \mu_W(G_j^B) = 1 - \tau$                     |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B > G_{j,y}^A$                 | $\tau > 1$            | $\mu_{SW}(G_j^B) = \frac{1}{\tau}, \mu_S(G_j^B) = 1 - \frac{1}{\tau}$ |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B > G_{j,y}^A$                 | $0 < \tau \leq 1$     | $\mu_{NW}(G_j^B) = \tau, \mu_W(G_j^B) = 1 - \tau$                     |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B < G_{j,y}^A$                 | $\tau > 1$            | $\mu_{NW}(G_j^B) = \frac{1}{\tau}, \mu_N(G_j^B) = 1 - \frac{1}{\tau}$ |
| $G_{j,x}^B = G_{j,x}^A, G_{j,y}^B > G_{j,y}^A$                 | $\tau$ does not exist | $\mu_N(G_j^B) = 1$                                                    |
| $G_{j,x}^B = G_{j,x}^A, G_{j,y}^B < G_{j,y}^A$                 | $\tau$ does not exist | $\mu_S(G_j^B) = 1$                                                    |
| $G_{j,x}^B > G_{j,x}^A, G_{j,y}^B = G_{j,y}^A$                 | $\tau = 0$            | $\mu_E(G_j^B) = 1$                                                    |
| $G_{j,x}^B < G_{j,x}^A, G_{j,y}^B = G_{j,y}^A$                 | $\tau = 0$            | $\mu_W(G_j^B) = 1$                                                    |
| $G_{j,x}^B = G_{j,x}^A, G_{j,y}^B = G_{j,y}^A$                 | $\tau$ does not exist | $\mu_C(G_j^B) = 1$                                                    |

## References

- Gao, S. A review of recent researches and reflections on geospatial artificial intelligence. *Geomat. Inf. Sci. Wuhan Univ.* **2020**, *45*, 1865–1874.
- Li, J.; He, Z.; Plaza, J.; Li, S.; Chen, J.; Wu, H.; Wang, Y.; Liu, Y. Social media: New perspectives to improve remote sensing for emergency response. *Proc. IEEE* **2017**, *105*, 1900–1912. [\[CrossRef\]](#)
- Zheng, N.; Guo, D. A spatial scene reconstruction framework in emergency response scenario. *J. Saf. Sci. Resil.* **2024**, *5*, 400–412. [\[CrossRef\]](#)
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv* **2023**, arXiv:2307.01952.
- Baraheem, S.S.; Nguyen, T.V. Sketch-to-image synthesis via semantic masks. *Multimed. Tools Appl.* **2024**, *83*, 29047–29066. [\[CrossRef\]](#)
- Maheshwari, P.; Chaudhry, R.; Vinay, V. Scene graph embeddings using relative similarity supervision. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2328–2336.
- Johnson, J.; Gupta, A.; Fei-Fei, L. Image generation from scene graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1219–1228.

8. Ashual, O.; Wolf, L. Specifying object attributes and relations in interactive scene generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4561–4569.
9. Vo, D.M.; Sugimoto, A. Visual-relation conscious image generation from structured-text. In Proceedings of the European Conference on Computer Vision, Virtual, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 290–306.
10. Yang, L.; Huang, Z.; Song, Y.; Hong, S.; Li, G.; Zhang, W.; Cui, B.; Ghanem, B.; Yang, M.H. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv* **2022**, arXiv:2211.11138.
11. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10684–10695.
12. Lian, L.; Li, B.; Yala, A.; Darrell, T. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv* **2023**, arXiv:2305.13655.
13. Wang, R.; Chen, Z.; Chen, C.; Ma, J.; Lu, H.; Lin, X. Compositional text-to-image synthesis with attention map control of diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5544–5552.
14. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. Pmlr, Virtual, 18–24 July 2021; pp. 8748–8763.
15. Feng, W.; He, X.; Fu, T.J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X.E.; Wang, W.Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv* **2022**, arXiv:2212.05032.
16. Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph. (TOG)* **2023**, *42*, 1–10. [[CrossRef](#)]
17. Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; et al. Reco: Region-controlled text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 14246–14255.
18. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 3836–3847.
19. Li, Y.; Wang, H.; Jin, Q.; Hu, J.; Chemerys, P.; Fu, Y.; Wang, Y.; Tulyakov, S.; Ren, J. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Adv. Neural Inf. Process. Syst.* **2024**, *36*. [[CrossRef](#)]
20. Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 1755–1764.
21. Li, D.; Li, J.; Hoi, S. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Adv. Neural Inf. Process. Syst.* **2024**, *36*. [[CrossRef](#)]
22. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
23. Liu, X.; Park, D.H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; Darrell, T. More control for free! image synthesis with semantic diffusion guidance. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 289–299.
24. Kavar, B.; Ganz, R.; Elad, M. Enhancing diffusion-based image synthesis with robust classifier guidance. *arXiv* **2022**, arXiv:2208.08664.
25. Shenoy, R.; Pan, Z.; Balakrishnan, K.; Cheng, Q.; Jeon, Y.; Yang, H.; Kim, J. Gradient-Free Classifier Guidance for Diffusion Model Sampling. *arXiv* **2024**, arXiv:2411.15393.
26. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2022**, arXiv:2207.12598.
27. Yang, B.; Luo, Y.; Chen, Z.; Wang, G.; Liang, X.; Lin, L. Law-diffusion: Complex scene generation by diffusion with layouts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 22669–22679.
28. Baykal, G.; Karagoz, H.F.; Binhuraib, T.; Unal, G. ProtoDiffusion: Classifier-free diffusion guidance with prototype learning. In Proceedings of the Asian Conference on Machine Learning, PMLR, Hanoi, Vietnam, 5–7 December 2024; pp. 106–120.
29. Randell, D.A.; Cui, Z.; Cohn, A.G. A spatial logic based on regions and connection. *KR* **1992**, *92*, 165–176.
30. Prim, R.C. Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* **1957**, *36*, 1389–1401. [[CrossRef](#)]
31. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]
32. Sara, U.; Akter, M.; Uddin, M.S. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study. *J. Comput. Commun.* **2019**, *7*, 8–18. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.