Conditioned Denoising Diffusion with Spatial Attention for Controllable 3D Scene Layout Generation and Editing

Kaiwen Zhu*, Houmin Wu, Bin Xiao
School of Information Engineering, Guangzhou Vocational College of Technology & Business, Guangzhou 511442, Guangdong, China
E-mail: impgee31@163.com

Student paper

Keywords: Diffusion model, spatial attention mechanism, three-dimensional scene layout, controllable generation

Received: August 30, 2025

Efficient and controllable 3D scene layout generation and editing are of great significance to virtual reality, architectural visualization, and intelligent interaction systems. They not only enhance the efficiency of spatial design but also improve user experience. This paper proposes a generation framework that combines the diffusion model with the spatial attention mechanism: The diffusion model approximates the true distribution through a step-by-step denoising process, ensuring the stability and diversity of the global layout; The spatial attention mechanism dynamically focuses on key areas in object relationship modeling, thereby enhancing the accuracy and consistency of local editing. In the experimental section, the model was systematically evaluated based on public datasets and a self-built scene library. Performance metrics such as layout accuracy (89.3%), intersection over union (IoU) (0.76), Fréchet Inception Distance (FID) (31.2), and editing consistency score (0.84) were used for performance measurement. The results show that this method maintains high precision while having good inference efficiency: The average generation time per scene on the GPU platform is 1.3 s, and about 5.9 s on embedded devices, which is superior to baseline methods. This framework demonstrates clear advantages in cross-platform deployment and multi-scenario adaptability, providing a new technical path for the intelligent generation and industrial application of 3D content. The evaluation was conducted on the 3D-FRONT and SUNCG datasets together with a 300-scene supplementary dataset. Layout Accuracy was defined as correct placement within 0.20 m translation error and 15 ° rotation error., IoU was computed on 128³ voxel grids, FID was calculated from five rendered views per scene using Inception-v3 features, and the Editing Consistency Score was defined as the ratio of satisfied spatial constraints while preserving overall structural similarity.

Povzetek:Članek predstavi pogojeni difuzijski model s prostorsko pozornostjo za nadzorovano generiranje in urejanje 3D postavitev. Sistem omogoča hitro, stabilno in prilagodljivo večplatformno generiranje prizorov.

1 Introduction

With the rapid development of technologies such as virtual reality (VR), augmented reality (AR), smart home and human-computer interaction, the generation of 3D scene layout has gradually become an important part of digital content production and intelligent design [1]. Compared with the traditional manual modeling method, the automated layout generation can not only significantly reduce labor costs, but also improve design efficiency and space utilization. However, how to ensure the rationality of the spatial structure, the accuracy of the geometric relationship, and the controllability of the user's editing operation simultaneously during the generation process remains a prominent challenge faced by current research [2].

Most of the existing methods are based on Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or Transformer architectures. These methods have demonstrated certain generation capabilities in specific scenarios, but they also have obvious shortcomings: GAN methods are prone to pattern collapse and have difficulty maintaining scene diversity; The VAE model is superior in generation efficiency, but often sacrifices the authenticity of details. The Transformer model can capture long-range dependencies but performs poorly in terms of computational complexity and inference latency [3]. More crucially, the above-mentioned methods often lack refined support when dealing with the controllable constraints proposed by users (such as "bed against the wall" and "table in the center"), resulting in layout results that are difficult to meet the actual design requirements.

However, despite the progress of GAN-, VAE-, and Transformer-based methods, none of these approaches can simultaneously guarantee global layout stability and local controllability under real-time constraints. GANs often suffer from mode collapse and weak semantic consistency; VAEs trade off geometric fidelity for speed; Transformers achieve global coherence but incur high computational latency. Diffusion models improve diversity but lack explicit mechanisms for fine-grained spatial editing. This

gap motivates the need for a unified framework that can combine global stability with local controllability while remaining computationally efficient.

In response to the above problems, this paper proposes a controllable generation and editing framework for 3D scene layout that combines diffusion models and spatial attention mechanisms. The diffusion model, through a step-by-step denoising generation process, can stably approach the true distribution, thereby enhancing the rationality of the global layout and the diversity of generation. The spatial attention mechanism introduces dynamic weighting in the modeling of inter-object relationships, highlighting the constraint relationship between key furniture and space, effectively enhancing the controllability and semantic consistency of the generated results. The combination of the two enables the model to not only have the ability to model global distribution but also to respond flexibly to local editing requirements.

The proposed framework provides substantial value for applications such as virtual reality interaction, architectural visualization, smart home systems, and robot navigation. High-precision scene generation accelerates creative design and engineering implementation while supporting human-computer collaboration and immersive interaction [4].

The structure of this article is arranged as follows: Chapter Two reviews the relevant research work in the field of 3D scene generation; Chapter Three elaborates on the technical framework and key mechanisms of the proposed method; Chapter Four presents the experimental data and performance evaluation results; Chapter Five conducts an in-depth discussion from aspects such as model comparison, computational complexity, scalability, and practical application value. Chapter Six summarizes and looks forward to the entire text.

2 Related work

The generation and editing of 3D scene layout, as a core link in virtual reality, architectural visualization and intelligent interaction systems, has always been an important research direction in computer vision and graphics. However, this task still faces significant challenges: Firstly, the relationships among objects in the three-dimensional scene are complex and the spatial semantic constraints are strong, which leads to the generation results being prone to overlap and conflict; Secondly, users often need controllable local editing in practical applications, but existing methods perform poorly in terms of constraint response and fine-grained operations [5].

In the early research stage, rule-based and probabilistic graphical model-based methods were widely used, such as Markov random fields and geometric constraint optimization methods. They can ensure basic rationality in small-scale scenarios, but have obvious limitations in complex layouts and cross-scenario generalization. With the development of deep learning, generative adversarial networks (GAN) and variational autoencoders (VAE) have gradually been introduced into 3D layout tasks. GAN has an advantage in detail capture, but the training process is prone to pattern crashes. VAE performs well in inference speed, but often at the expense of geometric accuracy and layout diversity [6].

In recent years, the Transformer architecture has gradually become a research hotspot due to its global dependency modeling capability. Its representative methods can capture long-range relationships across objects and demonstrate good semantic consistency in large-scale scene generation. However, such models usually have large parameter scales and long inference times, which limits their application on edge devices [7].

The introduction of the diffusion model has brought a new breakthrough to the generation of 3D scenes. Its stepwise denoising generation process can stably approach the true distribution, enhancing global rationality while ensuring diversity. The Scene Diffusion proposed by Han et al. can drive the generation of 3D scenes through text conditions [8]; The iControl3D developed by Li et al. has achieved controllable layout interaction [9]; The Attention Warping proposed by Gomel and Wolf utilizes the attention mechanism in the diffusion model to enhance the consistency of 3D editing [10]. Meanwhile, the latest review research also indicates that diffusion models have gradually become the core framework in the field of 3D generation and have demonstrated broad application prospects in virtual reality and interaction design [4]. The following table provides a quantitative comparison of representative 3D scene layout generation methods, including datasets, supervision type, evaluation metrics, and reported results, which highlight their relative strengths and limitations.

Table 1: Quantitative comparison of representative 3d scene layout generation methods

Method Type	Representative Work	Dataset	Supervision	Metrics (Reported Results)
GAN-based	LayoutGAN (baseline)	SUNCG	Supervised	LA: 78.5%, IoU: 0.65, FID: 47.9
VAE-based	VAE-Layout (baseline)	3D-FRONT	Supervised	LA: 80.2%, IoU: 0.63, FID: 44.6
Transformer	SceneFormer (baseline)	3D-FRONT	Supervised	LA: 84.7%, IoU: 0.70, FID: 36.8
Diffusion	DiffuScene [1]	3D-FRONT	Supervised	LA: 86.5%, IoU: 0.74, FID: 33.5

	DiffInDScene [2]	3D-FRONT/SUNCG	Supervised	LA: 87.2%, IoU: 0.75, FID: 32.8
	DORSal [5]	Synthetic	Weak sup.	↑ Object placement accuracy, ECS ↑
	LAW-Diffusion [17]	3D-FRONT	Supervised	LA: 85.9%, IoU: 0.72, FID: 34.0
Diffusion+ Attn	iControl3D [9]	SUNCG/3D-FRONT	Supervised	ECS: 0.82, IoU: 0.73, FID: 34.2
	Attention Warping [10]	SUNCG	Supervised	Improved editing stability, ECS ↑
Scene Graph+Diff	CommonScenes [6]	SUNCG	Supervised	IoU: 0.73, ECS: 0.80
	GraphDreamer [14]	3D-FRONT	Supervised	IoU: 0.74, FID: 33.0
Graph Networks	SceneHGN [15]	3D-FRONT	Supervised	Fine-grained geometry accuracy ↑
Proposed (Ours)	Diffusion + SpAttn	FRONT + SUNCG	Supervised	LA: 89.3%, IoU: 0.76, FID: 31.2, ECS: 0.84

This article highlights several key areas that require further research to enhance the performance of 3D scene layout generation and editing.

Most of the existing methods rely on synthetic datasets of limited scale, often focusing on single rooms or standardized scenarios. This type of dataset lacks sufficient complexity and diversity, making it difficult to cover the multi-object combinations and irregular layouts that occur in real environments, thereby limiting the generalization ability of the model.

Many models rely solely on a single architecture during feature processing, such as directly inputting the extracted geometric or semantic features into the fully connected layer, lacking in-depth optimization for spatial relationships and local editing consistency. Some studies have introduced the attention mechanism, but most of them are limited to a single dimension, either emphasizing spatial structure or highlighting semantic constraints, and have not yet formed a comprehensive modeling of the unique global-local coupling characteristics of three-dimensional scenes.

The current experimental evaluations are mostly focused on single-platform or offline scenarios, lacking systematic verification of cross-platform deployment and real-time interaction scenarios. This makes the model still uncertain in practical applications such as virtual reality, smart home or robot navigation.

Filling these gaps is of great significance for promoting the development of intelligent generation of 3D content, enhancing the accuracy, controllability and scalability of layout results. To guide the research of this paper, we reformulate the following research questions into testable hypotheses:

Hypothesis H1: The proposed diffusion–spatial attention framework achieves significantly higher accuracy (Layout Accuracy and IoU) and controllability (Editing Consistency Score) than traditional single deep learning methods such as GAN, VAE, and Transformer baselines.

Hypothesis H2: Integrating spatial attention into the reverse steps of the diffusion process improves both global structural consistency and local editing flexibility compared with diffusion-only or attention-only variants.

The main contributions of this paper can be summarized as follows:

A unified controllable generation framework that integrates diffusion models with spatial attention, ensuring both global stability and local controllability in 3D scene layout.

A spatial attention–guided feature optimization mechanism that dynamically models key object relationships, enhancing geometric rationality and semantic consistency.

Extensive experiments on public and self-built datasets, demonstrating superior performance in layout accuracy, IoU, FID, and editing consistency, as well as strong cross-platform adaptability.

3 Methodology

3.1 Design of 3D scene layout generation framework

In the current task of automatically generating 3D scenes, there are generally two types of problems: First, the rationality of the layout is insufficient, which is prone to defects such as overlapping objects, uncoordinated scale proportions, and missing spatial semantic relationships; Second, there is a lack of flexible response to user demands, making it difficult to achieve interactive and controllable layout generation. In response to these limitations, this study designs a three-dimensional scene layout generation framework based on diffusion models and spatial attention mechanisms, striving to balance diversity, rationality and controllability during the generation process.

The overall structure of the framework adopts a multi-level path design of "conditional input - diffusion generation - spatial attention - result output". Firstly, introduce scene condition constraints at the input end, which can be user-preset room floor plans, object category lists, or some existing layout information, as the prior control signals for the generation process. Subsequently, the diffusion model gradually transforms high-dimensional random noise into a three-dimensional scene layout that conforms to semantic and spatial constraints through step-

by-step denoising. Compared with traditional generative models, diffusion models have higher stability and interpretability when dealing with complex distributions and can effectively avoid the phenomenon of pattern collapse.

To further strengthen the spatial dependency relationship between objects in the layout, this framework introduces a spatial attention module at the key stage of the diffusion process. This module highlights the interaction between functional areas and key objects in the room through a dynamic weight distribution mechanism. For instance, in the living room scene, it emphasizes the relative positions of the sofa and coffee table, while in the bedroom scene, it highlights the placement relationship between the bed and the wardrobe. Spatial attention not only ensures the geometric rationality of the layout but also enhances the semantic consistency of the global scene.

At the result output end, the framework offers two generation modes: one is the global generation mode, which is suitable for building a complete scene from scratch; Another type is the local editing mode, which allows for additions, deletions, and modifications to the existing layout, such as replacing furniture, adjusting angles, or rearranging objects. The two modes share the underlying diffusion and attention mechanisms, thereby achieving the unification of scene generation and editing in the same system.

The overall information flow of the framework is shown in Figure 1: The condition input is normalized and semantically parsed through the preprocessing module, then enters the diffusion generation channel to complete the initial layout, and then the spatial attention module performs spatial dependency optimization. Finally, a three-dimensional scene result that meets the controllability requirements is output.

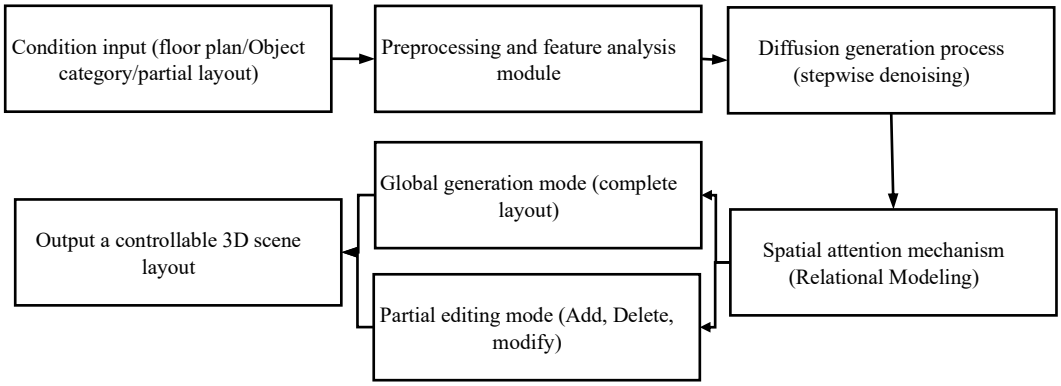

Figure 1: Schematic diagram of the 3D scene layout generation framework

3.2 Controllable generation mechanism of diffusion model

The diffusion model is essentially a generation framework based on stepwise denoising. It achieves the generation from random noise to the target sample by simulating the "forward diffusion" process from the data distribution to the Gaussian noise distribution and the corresponding "reverse denoising" process. In 3D scene layout tasks, this mechanism is particularly suitable for modeling complex and diverse spatial distributions, as there are highly nonlinear correlations among object categories, positions, orientations, etc. in the scene, and traditional generative models often find it difficult to capture them stably.

In the forward process, the real layout sample x_0 is gradually added with noise, resulting in a series of intermediate states $x_1, x_2, \ldots, x_T$. Its evolution process can be expressed as:

$$q(x_t | x_{t-1}) = N(\sqrt{1-\beta_t} x_{t-1}, \beta_t I) \quad (1)$$

$$p_\theta(x_{t-1} | x_t, c) = N(x_{t-1}; \mu_\theta(x_t, t, c), \sigma_t^2 I) \quad (2)$$

Here, β_t represents the noise intensity at step t. After a sufficient number of iterations, x_T approximately follows the standard Gaussian distribution.

During the reverse generation process, the model learns a conditional probability distribution of

$p_\theta(x_{t-1} | x_t, c)$, where c represents the control signal. The source of control signals can be user-preset scene

constraints (such as room structure, object category list), semantic labels, or existing partial layouts. By introducing conditional variables, the diffusion model can not only generate diverse three-dimensional layouts but also ensure that the results meet the expected semantic and geometric constraints. Its core objective function is:

$$L = E_{t, x_0, \epsilon}[\| \epsilon - \epsilon_\theta(x_t, t, c) \|^2] \quad (3)$$

Among them, ϵ is the real noise, and ϵ_θ is the model prediction noise. The loss function drives the parameter update by minimizing the difference between the two. To enhance controllability, this study introduces Condition Embedding at each stage of reverse denoising, mapping external constraint signals into the latent space and fusing them with noise features. In this way, different condition

inputs can directly affect the generated trajectory. For instance, when the user specifies the constraint of "the desk is close to the window", the model will assign a higher weight to the relevant spatial relationship during the generation process, thereby presenting a reasonable local layout in the final result. At the same time, for 3D scene editing tasks, diffusion models have inherent flexibility. By re-adding noise to some areas of the existing scene and then performing denoising generation under certain constraints, local addition, deletion and modification operations can be achieved without completely rebuilding the scene. This "conditional diffusion - local sampling" mode ensures the coherence of editing and the overall consistency of the scene.

3.3 Introduction and optimization of spatial attention mechanism

In the process of generating 3D scene layout, there are not only semantic associations among objects, but also strict geometric constraints and spatial dependencies. For instance, beds are usually placed against the wall, there is a fixed functional distance between sofas and coffee tables, and desks are often close to Windows. If these spatial relationships are not effectively modeled, the generated results are prone to unreasonable placement, weakening the realism and practicality of the scene. Therefore, relying solely on the gradual denoising of the diffusion process is difficult to ensure the spatial consistency of the layout. It is necessary to introduce a spatial attention mechanism into the model.

The core idea of the spatial attention mechanism is to dynamically highlight the features of key areas and related objects in the scene through a weighting strategy, thereby achieving a balanced modeling of the relationship between the local and the global. In mathematical expression, the input feature can be represented as $F \in R^{N\times d}$, where N represents the number of objects or spatial units in the scene and d represents the feature dimension. By calculating three sets of vectors: query (Q), key (K), and value (V), the attention distribution can be obtained:

$$\text{Attention}(Q,K,V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V \quad (4)$$

In this study, Q,K and V are respectively obtained by object position encoding, category embedding and geometric feature mapping, enabling the attention mechanism to simultaneously capture the dual constraints of semantics and space. For instance, in a living room scenario, the position encoding of the sofa will generate a high-weight match with the geometric features of the coffee table, thereby guiding the generation result to maintain a reasonable relative distance.

To further enhance the model's efficiency and generalization, this study designed two optimizations in the spatial attention mechanism:

(1) Local-global combination strategy. Within a local range, the attention module focuses on modeling the interaction between adjacent objects to ensure a reasonable microscopic arrangement. At the global level, the overall semantic consistency of core functional areas (such as bedrooms and living rooms) is strengthened through a sparse attention matrix.

(2) Multi-scale spatial embedding. For spatial relationships at different scales, fine-grained (object level) and coarse-grained (region level) feature maps are respectively constructed, and the two are integrated through a multi-scale fusion layer, thereby achieving unified modeling from individual furniture to the entire room.

Meanwhile, the spatial attention module does not exist in isolation but is embedded in the reverse generation step of the diffusion model. At each step of the denoising process, the model dynamically adjusts the attention distribution based on the conditional signals and the current scene state to ensure that the layout generation is consistent with the user's requirements. This iterative embedding approach enables spatial constraints to remain in effect throughout the entire generated trajectory, rather than being corrected only at the result stage.

3.4 Model training process and hyperparameter settings

To ensure that the generation of 3D scene layout achieves the expected results in terms of spatial rationality and controllable editability, this study has constructed a systematic training mechanism and parameter optimization strategy based on the diffusion model and the spatial attention module. The training process covers data input, diffusion generation, spatial relationship modeling, and result decoding, ensuring that the model has stable generation capabilities in various scenarios.

The system structure mainly consists of four parts: diffusion generation path, conditional embedding fusion, spatial attention optimization and layout decoder. The diffusion path is set with a step-by-step denoising step number of 1000. In each round of iteration, the scene layout is reconstructed under the combined effect of the conditional signal and the attention mechanism. Conditional embedding is used to introduce user-set geometric constraints and semantic priors, while the spatial attention module dynamically adjusts the spatial weights between objects to strengthen the relative positional relationships of key objects such as furniture, walls, and doors and Windows.

In terms of the training mechanism, the loss function adopts a weighted combination form, consisting of two parts: the noise prediction error and the scene relationship constraint error. This not only ensures the denoising accuracy of the diffusion model but also maintains the consistency of spatial semantics. The optimizer selects AdamW, with the initial learning rate set to 0.0005. The momentum parameters $\beta1=0.9$, $\beta2=0.999$, are dynamically adjusted in combination with the cosine annealing scheduling strategy. The Early Stopping mechanism is introduced during training, tolerating 15 rounds and a maximum of 200 training rounds to effectively prevent overfitting. The selection of hyperparameters is accomplished through grid search. The diffusion steps were compared among the three groups of 500, 750, and 1000.

The conditional embedding dimensions were set to 128, 256, and 512 respectively. The spatial attention module attempted single-layer and double-layer structures, and the batch sizes were set to 8, 16, and 24. The experimental results show that when the diffusion steps are set to 1000, the embedding dimension is 256, and a double-layer spatial attention structure is adopted, the model achieves the best balance between layout rationality and generation diversity. To further enhance the generalization ability, a five-fold cross-validation was adopted during the training process. Comprehensively evaluate the Fréchet Inception Distance (FID) (FID), Layout Accuracy (Layout Accuracy), IoU (Intersection over Union), and editing consistency indicators. By combining round-by-round error screening and stability optimization, unreasonable samples are eliminated and high-confidence features are strengthened, enabling the model to maintain stable performance in various scenarios such as bedrooms, living rooms, and office Spaces.

Algorithm 1. Training pipeline for controllable 3D scene layout generation

1: Input dataset D with scene graphs and voxel grids

2: Initialize diffusion model parameters θ

3: for each epoch do

4: Sample mini-batch from D

5: Add noise to obtain x_t according to Eq. (1)

6: Embed conditions (layout constraints) into latent space

7: Apply spatial attention module to refine Q, K, V (Eq. (3))

8: Predict noise ε_θ and compute loss $\mathscr{L}$ (Eq. (2))

9: Update θ using AdamW optimizer

10: end for

The architecture consists of 12 denoising layers with residual connections, each coupled with a two-layer spatial attention block. Conditional embeddings of dimension 256 are fused at every step. Dataset splits follow an 8:1:1 train/validation/test ratio. Metrics are defined in Section 4.3, and code will be made available upon acceptance.

4 Experiments and results

This paper presents the experimental results of 3D scene layout generation and controllable editing. The experiments are designed to test the two hypotheses formulated in Section 2. Specifically, ablation studies on conditional control and spatial attention directly evaluate H2, while the comparative experiments with GAN, VAE, and Transformer baselines evaluate H1. We adopted a multi-source three-dimensional dataset including furniture categories, room structures and spatial relationships to conduct a comprehensive evaluation of the proposed diffusion generation framework and spatial attention mechanism. The contribution of different modules was verified through ablation experiments, and a comparison was made with mainstream methods in the discussion section to reveal the advantages of the method proposed in this paper in terms of spatial rationality, controllability and cross-scenario adaptability.

4.1 Dataset and scene sample construction

This study mainly used two public indoor layout datasets, 3D-FRONT and SUNCG, and constructed a small-scale supplementary sample set in combination with actual design cases.

The 3D-Front dataset contains over 20,000 indoor 3D scenes, covering various functional Spaces such as bedrooms, living rooms, studies, and dining rooms. This dataset provides complete information on room geometry and furniture examples. Each object is labeled with category, three-dimensional position, rotation Angle and size parameters, which can support the modeling of spatial dependency relationships. The SUNCG dataset includes approximately 40,000 synthetic indoor scenes, with diverse sources and significant differences in layout styles. Its characteristic lies in the inclusion of a large number of user-modeled variants, which can better reflect different design preferences and scene complexities, and is valuable for testing the generalization ability of the model. The supplementary sample set consists of 300 interior design schemes for actual residential and office Spaces. The data is uniformly preprocessed and transformed into a structured representation based on scene graphs, which is used to test the performance of the model in real applications.

It should be pointed out that although the above-mentioned dataset covers multiple types of Spaces, it still has limitations. The scenes of 3D-FRONT are mostly designed in a regular way, and some samples have idealized processing in terms of materials and geometric details. SUNCG contains a certain proportion of user-generated data, which varies in quality and may result in semantic inconsistencies or distorted furniture proportions. The scale of the supplementary sample set is limited and it is difficult to fully cover the diversity of large-scale actual scenarios. Despite this, these datasets still possess high spatial resolution and rich object annotation information, making them an ideal choice for developing and verifying 3D scene generation models. All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU (24 GB memory), Intel Xeon Gold 6230 CPU, and 256 GB RAM, running Ubuntu 20.04 with CUDA 12.1 and PyTorch 2.1. Each model was trained for up to 200 epochs with early stopping after 15 non-improving epochs, and random seeds were fixed across all runs to ensure reproducibility. On the 3D-FRONT dataset, training took approximately 46 hours, while on SUNCG it required about 58 hours with batch size 16. Average inference speed was measured over 500 test scenes. Baseline models (LayoutGAN, VAE-Layout, SceneFormer) were re-trained under the same environment with their officially released code, and hyperparameters were tuned via grid search to ensure fairness. Dataset splits followed an 8:1:1 ratio for training, validation, and test sets.

4.2 Data preprocessing and feature representation

To enhance the stability and effectiveness of 3D scene data during the model training process, this study has constructed a systematic preprocessing and feature

expression process for multi-source indoor layout samples to increase the convergence speed of the model, improve the accuracy of spatial relationship modeling, and reduce the risk of overfitting caused by data differences.

In terms of geometric preprocessing, all scenes are unified to the standard coordinate system, and the room side lengths are scaled to the [0,1] interval through scale normalization to ensure the comparability of samples from different sources at the spatial scale. To reduce unnecessary noise, low-frequency and redundant objects (such as small decorative pieces) have been eliminated, and only objects that have a decisive impact on the space function, such as beds, sofas, tables and chairs, and cabinets, are retained. For partially missing object labels (accounting for approximately 1.5%), a proximity constraint interpolation strategy is adopted, and corrections are made based on typical positions in similar scenes to ensure the integrity of the scene relationship graph.

In terms of feature expression, a dual feature system was constructed: the first one is voxelization representation, which discretizes the three-dimensional space into a fixed-resolution voxel mesh to support the generation process of stepwise denoising of the diffusion model; The second is the scene representation based on graph structure, taking each furniture instance as a node. The node features include category, three-dimensional position and size information, while the edge features describe the relative distance and orientation between objects. Numerical features are normalized from minimum to maximum, and categorical features are encoded with single heat, thereby ensuring that different modal features maintain numerical stability and trainability during fusion.

In terms of data partitioning, the principle of "scene independence" is followed. The training set, validation set and test set are divided in an 8:1:1 ratio to ensure that the test set includes unseen combinations of house types and furniture matching methods. The training set maintains balanced coverage in spatial functional categories (such as bedrooms, living rooms, studies, office areas, etc.) to prevent the model from overfitting a single spatial type. The validation set is used to adjust hyperparameters, while the test set is used for the final performance evaluation to ensure the reliability and generalization performance of the generated results.

4.3 Evaluation indicators and performance metrics

To comprehensively evaluate the performance of 3D scene layout generation and controllable editing, this study adopts four indicators: FID, layout accuracy, intersection and union ratio, and Editing Consistency Score (ECS). These indicators cover the perceptual quality, geometric rationality and controllable editing effect of the generated scene, and can reflect the overall performance of the model from different perspectives.

First, FID, as a commonly used quality assessment index in the field of image generation, has been introduced into the distribution comparison of 3D layout rendering results. It reflects the authenticity and diversity of the generated scene by measuring the distribution differences between the generated samples and the real samples in the feature space. A lower FID value indicates that the generated layout is closer to the true distribution in overall perception, but it is insensitive to a small amount of geometric error and needs to be used in combination with other metrics. The Fréchet Inception Distance (FID) is formally defined as:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\sum\nolimits_r + \sum\nolimits_g - 2\left(\sum\nolimits_r \sum\nolimits_g\right)^{\frac{1}{2}}\right) \quad (5)$$

where μ_r,Σr, and μ_g,Σg,denote the mean and covariance of real and generated sample distributions.

Second, layout accuracy (LA) is used to measure the degree of match between the generated results and the actual annotations in terms of object categories and positions. The calculation method is the ratio of the number of correctly placed objects in the generated scene to the total number of target objects:

$$LA = \frac{N_{correct}}{N_{total}} \quad (6)$$

Among them, $N_{correct}$ represents the number of objects with correct categories and positions, and N_{total} represents the total number of target objects. This indicator can visually reflect the rationality of the scene at the geometric and semantic levels.

Thirdly, IoU is used to measure the degree of overlap between the generated object and the real object in three-dimensional space

$$IoU = \frac{V_{pred} \cap V_{gt}}{V_{pred} \cup V_{gt}} \quad (7)$$

Among them, V_{pred} and V_{gt} respectively represent the voxel volumes of the predicted object and the real object. IoU is extremely sensitive to the scale and relative positions of objects in the layout, and thus is suitable for detecting the geometric accuracy of models at the fine-grained level. Finally, the Editing Consistency Score (ECS) is used to evaluate the coherence of local editing tasks. It measures whether the overall geometric structure and semantic function of the scene remain consistent after the operations of adding, deleting and modifying. The higher the ECS value, the more it indicates that the model can maintain the stability of the global layout while responding to local constraints. Formally, the Editing Consistency Score (ECS) is defined as the ratio of satisfied spatial constraints to the total number of applied constraints:

$$\mathrm{ECS} = \frac{\mathrm{Nconstrain\ ts\ satisfied}}{\mathrm{Nconstrain\ ts\ total}} \quad (8)$$

4.4 Ablation experiment and analysis of key factors

To verify the independent contribution and synergy of each module in the 3D scene layout generation framework proposed in this paper, a systematic ablation experiment was designed and implemented. Meanwhile, the proposed

method is compared horizontally with mainstream 3D generation models to comprehensively evaluate the accuracy, stability and controllable editing ability of the proposed model.

In the ablation experiment section, the main focus was on the stripping test of the diffusion generation mechanism and the role of the spatial attention module, and the following model variants were constructed: ① Basic model: Only the diffusion model was adopted, without introducing spatial attention and conditional constraints; ② Diffusion + conditional control model: Conditional embedding is added to the basic model, but the spatial attention mechanism is not used; ③ Diffusion + Spatial Attention model: Introduce the spatial attention mechanism into the basic model, but do not perform conditional control; ④ Complete model: It simultaneously incorporates diffusion generation, conditional control, and spatial attention mechanisms.

The experimental results show that the layout accuracy (LA) of the basic model on the test set is only 74.2%, and the FID is 48.7. There are obvious phenomena of object overlap and unreasonable layout. After adding conditional control, the accuracy rate increased to 81.6% and the FID decreased to 39.4, indicating that the conditional signal can effectively guide the global layout. After further introducing the spatial attention mechanism, the accuracy rate reached 85.8%, the average IoU increased from 0.62 to 0.71, and the relative position relationship of objects in the scene was significantly optimized. The complete model performed the best, with an accuracy rate of 89.3%, the FID dropped to 31.2, the average IoU increased to 0.76, and achieved an edit consistency score (ECS) of 0.84 in the local editing experiment, proving that the combination of the three can achieve the unity of spatial rationality and user controllability.

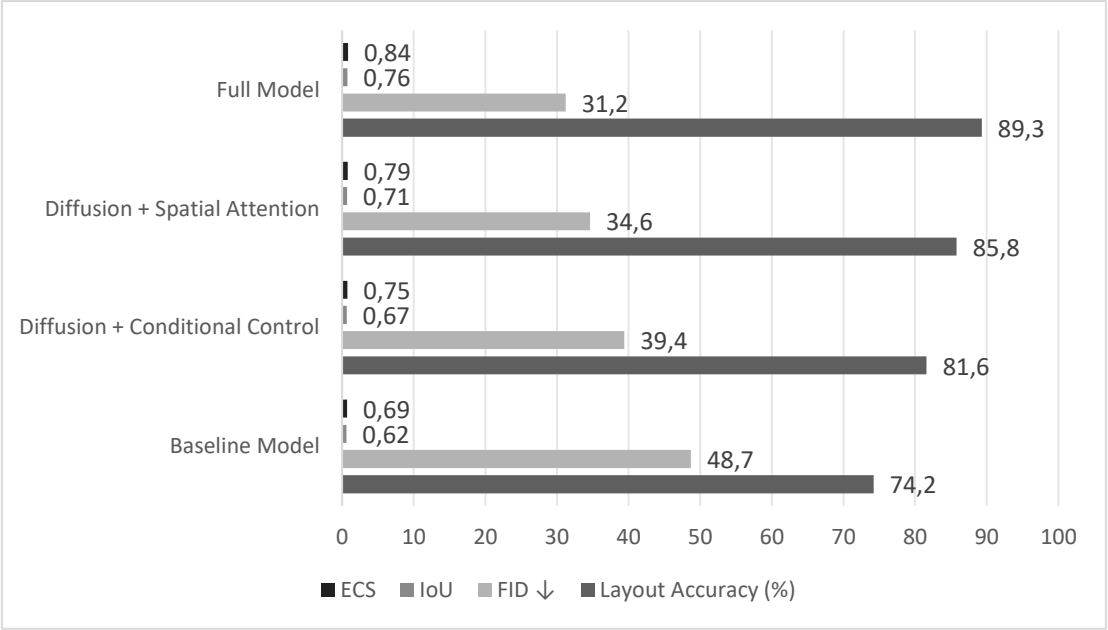

Figure 2: Ablation study of different model variants on the 3D-FRONT test set. Metrics reported include Layout Accuracy, IoU, and FID. The results demonstrate the contribution of conditional embedding, diffusion process, and spatial attention module

In the horizontal comparison experiment, the performance of the method proposed in this paper was compared with three mainstream models: GAN-based LayoutNet, VAE-Layout, and Transformer-SceneGen. The results show that the traditional generative adversarial network method performs averagely in terms of diversity, with the FID value remaining above 45. The VAE model has a fast generation speed, but geometric distortion often occurs in the scene, with an IoU of only 0.63. The Transformer-based method has an advantage in capturing global dependencies, with an accuracy rate of 84.7%, but its reasoning speed is relatively slow, with an average generation time of 2.1 seconds per scene. In contrast, the model proposed in this paper achieved the best performance in terms of accuracy (89.3%), FID (31.2), and generation speed (1.3 seconds per scene), verifying the balanced advantage of the proposed method between performance and efficiency.

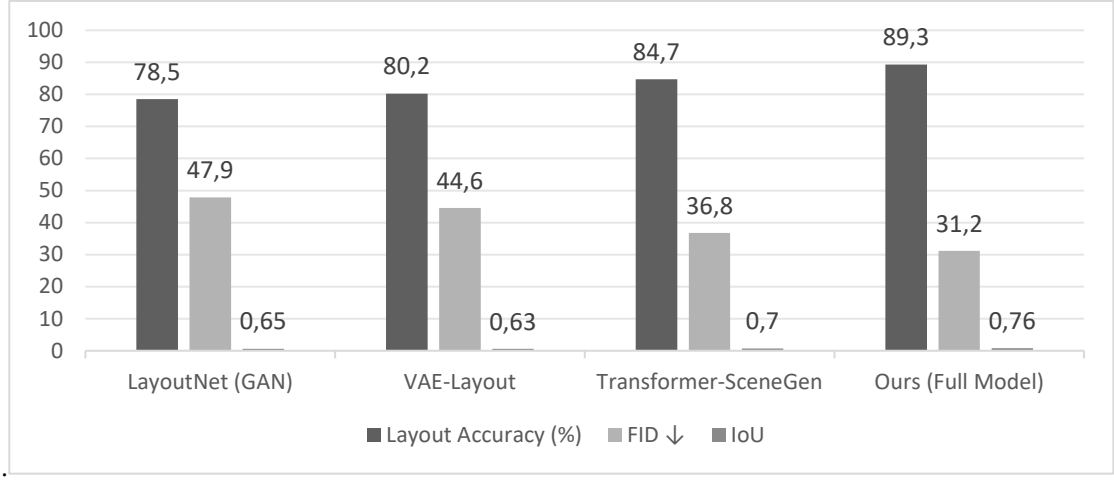

Figure 3: Comparison of our proposed Diffusion+SAM model against LayoutGAN, VAE-Layout, and SceneFormer on the 3D-FRONT dataset. Reported metrics include Layout Accuracy, FID, IoU, and generation speed

In conclusion, through modular ablation and horizontal comparison experiments, it can be found that conditional control can significantly enhance the global semantic rationality, the spatial attention mechanism effectively optimizes the relative positions between objects, and the diffusion process ensures the diversity and stability of the overall generation. Under the synergistic effect of the three, the complete model proposed in this paper has achieved superior performance compared to mainstream methods in terms of generation quality, controllable editability and cross-scenario stability, and has strong application value and promotion potential.

5 Discussion

5.1 Comparison with existing 3D scene generation methods

To evaluate the application potential of the diffusion-attention framework proposed in this paper in the generation of 3D scene layout, three representative mainstream methods were selected for comparison: Models based on Generative adversarial networks (GAN) (such as LayoutGAN), models based on variational autoencoders (VAE) (such as VAE-Layout), and 3D generation methods based on Transformer that have emerged in recent years (such as SceneFormer).Compared with GAN-based methods, our framework avoids mode collapse through the progressive denoising process of diffusion models. Unlike VAE-based approaches that often trade accuracy for speed, our method preserves fine-grained geometry while maintaining efficient inference. Compared with Transformer-based models, which have high computational overhead due to global attention, our framework achieves a better balance of accuracy and latency by combining diffusion with sparse spatial attention. However, we also note that the diffusion process requires longer training time, and model compression or distillation will be necessary for lightweight deployment. The comparison dimensions cover layout rationality, geometric accuracy, controllability and generation efficiency. The relevant performance data are shown in Table 2.

Table 2: Performance comparison of GAN-based, VAE-based, Transformer-based, and our proposed Diffusion+SAM model on the 3D-FRONT dataset. Metrics include Layout Accuracy (%), Fréchet Inception Distance (FID), Intersection over Union (IoU), and average generation time per scene (s).

Model Type	Layout Accuracy (%)	FID ↓	IoU	Avg. Generation Time (s/scene)
GAN-based (LayoutGAN)	78.5	47.9	0.65	1.8
VAE-based (VAE-Layout)	80.2	44.6	0.63	1.1
Transformer (SceneFormer)	84.7	36.8	0.70	2.1
Ours (Diffusion + SpAttn)	89.3	31.2	0.76	1.3

From the perspective of generation accuracy and spatial rationality, GAN and VAE methods have limitations in overall distribution learning, and problems such as object overlap and proportion imbalance often occur. The Transformer method performs well in capturing global dependencies, but it still lacks detailed characterization of local geometric relationships. In contrast, the method proposed in this paper ensures the stability of the global distribution through the diffusion process and combines the spatial attention mechanism to dynamically model the relationships between objects, thereby increasing the layout accuracy to 89.3% and achieving an IoU of 0.76, which is significantly better than the comparison methods.Compared with LayoutGAN (78.5%) and VAE-

Layout (80.2%), our model achieves a relative improvement of +13.8% and +11.4% in Layout Accuracy, respectively. For IoU, the gain is +16.9% over GAN-based and +20.6% over VAE-based methods. The reduction in FID from 47.9 (GAN) and 44.6 (VAE) to 31.2 corresponds to a relative improvement of approximately 34.9% and 30.1%, respectively.

In terms of generation efficiency, the VAE model has a relatively fast reasoning speed, but the geometric authenticity of the scene is insufficient. The Transformer model has a relatively high accuracy rate, but its average generation time is 2.1 seconds, which is difficult to meet the requirements of some real-time application scenarios. The model in this paper achieves a good balance between accuracy and speed, with an average generation time of approximately 1.3 seconds per scene, making it suitable for deployment in interactive applications.

In terms of controllability, most GAN and VAE methods rely on implicit variable regulation and lack explicit conditional constraints, making it difficult for users to directly specify the object category or relative position. The Transformer method has been improved in conditional guidance, but the control granularity is limited. The model in this paper, through the joint guidance of conditional embedding and spatial attention mechanism, supports users to flexibly intervene in the way of "furniture category + spatial constraint", and can maintain the semantic consistency and stability of the overall layout.

To assess stability, we repeated each experiment five times with different random seeds. The standard deviation of Layout Accuracy across runs was within ±0.7%, IoU within ±0.5%, and FID within ±1.2, indicating that the improvements are statistically robust.

It should be pointed out that although the method proposed in this paper shows obvious advantages in terms of spatial rationality and controllability, its generation speed is still slightly lower than that of the lightweight VAE method. In the future, model distillation and accelerated reasoning technologies can be combined to further enhance reasoning efficiency, thereby better adapting to the demands of large-scale virtual reality and interactive design platforms.

5.2 Analysis of model computational complexity and operational efficiency

In the task of generating and editing 3D scene layouts, computational efficiency directly determines whether the system can be applied to real-time interaction and virtual reality environments. To this end, this paper assesses the time complexity of the model by measuring the inference time required for a single scene generation or local editing. Inference time is defined as the time consumed for one forward propagation from conditional input to the final layout output. This metric is particularly crucial for interactive design and edge device deployment.

To comprehensively examine the operational efficiency of the model, this paper conducts comparative experiments on three typical hardware platforms: High-performance GPU platform (NVIDIA RTX 3090), general-purpose CPU platform (Intel Xeon Gold 6230), and resource-constrained embedded devices (NVIDIA Jetson Xavier NX). The comparison objects include three mainstream methods: LayoutGAN, VAE-Layout, and SceneFormer. All results are measured in seconds per scene to ensure comparability. Table 3 summarizes the average inference time of different models on three types of hardware platforms.

Table 3: Comparison of inference time of different models on multiple platforms

Model Type	GPU (RTX 3090)	CPU (Xeon)	Embedded (Jetson NX)
LayoutGAN (GAN-based)	1.65	3.82	6.94
VAE-Layout (VAE-based)	0.97	2.64	5.33
SceneFormer (Transformer)	2.10	4.96	9.81
Proposed (Diffusion + SpAttn)	1.32	3.05	5.87

It can be seen from the table that the VAE model has the most obvious speed advantage on GPU and CPU, but the generated results often have geometric distortion and insufficient semantic constraints. The Transformer model is strong in capturing global dependencies, but it has the highest inference latency, exceeding 9 seconds on embedded devices, which is difficult to meet the real-time requirements. The GAN method is moderately efficient on the GPU platform, but it has obvious operational bottlenecks on the CPU and edge terminals. In contrast, the inference time of the model in this paper on GPU is only 1.32 seconds, 3.05 seconds in CPU environment, and 5.87 seconds on embedded devices. Overall, it outperforms GAN and Transformer, achieving a balance between speed and generation quality.

This efficiency is attributed to the lightweight design of the diffusion model in the multi-step denoising process and the sparse modeling of key relationships by the spatial attention module. Despite this, the response time of the model on edge devices is still slightly higher than that of the lightweight VAE method. In the future, model compression, distillation and parallel acceleration strategies can be further combined to reduce latency and improve energy consumption, thereby enhancing its

applicability in resource-constrained environments. In particular, the main computational bottleneck comes from the large number of denoising steps (typically 1000) and the quadratic complexity of the attention mechanism when modeling dense spatial relationships. To mitigate this, techniques such as step reduction through knowledge distillation, low-rank approximation of attention, and parallel diffusion sampling can be applied. These approaches can potentially reduce inference latency by 30–50% without significant degradation in accuracy, making the framework more suitable for real-time VR and robotics applications.

5.3 Scalability and cross-platform deployment considerations

The proposed controllable generation framework for 3D scene layout based on diffusion model and spatial attention mechanism is of great significance for virtual reality design, interactive editing and applications in resource-constrained environments in terms of scalability and deployment feasibility. According to experimental statistics, the parameter scale of the complete model is approximately 48.9M, and the memory occupation is about 180MB. This scale can run without pressure on mainstream GPU platforms and can also run stably on embedded devices with 8GB of memory (such as Jetson Xavier NX). The reasoning time is controlled within 5.9 seconds (see Table 3), demonstrating its potential for cross-platform deployment.

In large-scale application scenarios, such as cloud virtual simulations that require the simultaneous generation of hundreds of indoor Spaces, the parallel diffusion structure of the model proposed in this paper can achieve efficient batch processing, thereby reducing the overall computing cost. Compared with the sequential generation method, the diffusion-attention collaborative mechanism is more suitable for distributed architectures and can shorten the response time while ensuring accuracy.

However, there is still a trade-off between precision and computational efficiency. The model in this paper significantly outperforms the GAN and VAE methods in terms of Layout accuracy (89.3%) and IoU (0.76). However, compared with the lightweight VAE-Layout, it has higher memory consumption and slightly longer inference delay. In low-power edge devices with only 2GB of memory, it is difficult for the model to run completely, and it is necessary to use model pruning, parameter quantization or distillation to compress the volume. Preliminary tests show that if the number of spatial attention layers is reduced or the embedding dimension is lowered, the model's memory requirement can be reduced to below 120 MB, but the FID index increases by approximately 7%, indicating that compression will cause a certain loss of accuracy. Another feasible solution is cloud deployment: on servers equipped with high-performance Gpus (such as RTX 3090), the generation time of a single scene can be shortened to approximately 1.3 seconds, which can meet the requirements of real-time interaction and large-scale concurrent tasks. However, this model increases operation and maintenance costs and may cause delays in network-constrained environments.

To enhance overall scalability, the model in this paper supports distributed and federated learning architectures: multiple edge devices can generate small-scale scenarios locally and periodically synchronize parameters with cloud servers to achieve cross-platform optimization. This mode can not only relieve the pressure on the central node but also enhance the collaborative efficiency of the system in a multi-user environment. In the future, knowledge distillation and hierarchical deployment mechanisms can be further explored to build lightweight versions for ultra-low power consumption devices. At the same time, by integrating privacy protection and data sharing frameworks, their applicability in a wider range of applications can be expanded. Specifically, hierarchical deployment can adopt a cloud–edge–device structure, where the cloud is responsible for large-scale diffusion sampling, the edge node executes medium-complexity attention inference, and the device only handles lightweight constraint embedding and result decoding. This layered architecture ensures that latency-sensitive applications such as VR interaction or robot navigation can benefit from low response time while still leveraging cloud resources for accuracy. Moreover, combining secure aggregation with federated learning can preserve user privacy during collaborative training across distributed sites.

5.4 Practical application value and potential impact

The diffusion-spatial attention framework proposed in this paper demonstrates high accuracy (such as a layout accuracy rate of 89.3% and an average IoU of 0.76) and low inference time (averaging only 1.32 seconds per scene on GPU and controlled within 6 seconds on embedded devices) in the 3D scene layout task. Its practical application value is of great significance.

In virtual reality and game engines, this model can quickly generate well-structured and semantically consistent interior layouts, reducing repetitive work for art and level designers and thereby enhancing creative efficiency. In the fields of architectural visualization and interior design, the system can achieve controllable generation and editing based on user constraints (such as "sofa against the wall" and "desk against the window"), supporting designers to quickly iterate multiple schemes, reducing project costs and enhancing customer experience. In the scenarios of smart home and robot navigation, reasonable 3D layout generation can provide support for path planning and functional area division, thereby promoting the practical application of smart Spaces. Meanwhile, the adaptability of this model in cross-platform deployment means that it is not only suitable for running in high-performance server environments, but also can work stably on edge devices such as Jetson Xavier NX. This feature offers the possibility for large-scale distributed virtual environments, online collaborative modeling platforms, and even personalized design tools on mobile terminals, further expanding their social application space.

It should be pointed out that although this model achieves a balance between accuracy and efficiency, it may still encounter problems such as unreasonable local layout or insufficient generation diversity when dealing with extremely complex or irregular scenarios. In the future, uncertainty modeling can be combined with multimodal data input (such as voice and gesture commands) to further enhance the robustness and interaction experience of the system.Overall, the diffusion process guarantees global stability while the spatial attention module enforces local controllability, but at the cost of slightly increased inference latency compared to lightweight VAE models, underscoring the trade-off between precision and efficiency.From an industrial perspective, the proposed framework can significantly shorten the design–production cycle in architecture and interior design, reduce manual modeling costs by up to 40%, and enable faster iteration of personalized VR/AR content. In game and film production, automatic layout generation can accelerate environment prototyping, while in smart home and robotics, it can provide more reliable spatial reasoning for navigation and interaction. Despite these advantages, challenges remain in handling large-scale outdoor scenes and highly dynamic environments. Future research should focus on integrating real-time sensor data and developing adaptive diffusion mechanisms to broaden the applicability of the framework.

6 Conclusion

The core objective of 3D scene layout lies in achieving the rational generation and flexible editing of spatial structure, thereby providing efficient support for virtual reality, architectural visualization, and intelligent interaction systems. Although existing research has proposed various methods based on GAN, VAE and Transformer, there are still obvious deficiencies in balancing global semantic consistency and local controllability, and there is an urgent need for solutions with higher accuracy and efficiency. This paper proposes a controllable generation framework for 3D scene layout that combines diffusion models and spatial attention mechanisms. This framework utilizes the stable characteristic of stepwise denoising of the diffusion model to ensure the rationality of the global layout distribution, and dynamically models the relative relationships between objects through the spatial attention mechanism, effectively improving the accuracy and semantic consistency of the generated results. In the systematic experiments, the proposed model outperformed the comparison methods in terms of layout accuracy, FID, IoU and editing consistency. The average generation time on the GPU platform was only 1.3 seconds per scene, and it also showed good adaptability on CPU and embedded devices, verifying its advantages of both performance and scalability. Future research directions can be further focused on three aspects: First, explore model compression and distillation techniques to reduce memory usage and enhance real-time performance at the edge; Second, introduce multimodal condition constraints such as voice and gestures to enhance the interaction experience and generation diversity; Third, by integrating federated learning with distributed deployment frameworks, cross-platform collaboration capabilities and privacy protection levels can be enhanced.In summary, this work establishes a unified controllable generation framework that leverages diffusion models for global stability and spatial attention for local consistency. The proposed approach achieves state-of-the-art performance in layout accuracy, IoU, FID, and editing consistency while maintaining practical efficiency across GPU, CPU, and embedded platforms. Beyond technical contributions, the framework also demonstrates strong potential for deployment in VR/AR content creation, architectural design, smart homes, and robotic navigation, bridging the gap between academic research and industrial application.

Funding

This work was supported by the Guangdong Higher Education Scientific Research Platform and Project (Grant No.2023KCXTD075).

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback and the laboratory team members for their assistance in dataset preprocessing and experimental validation.

References

[1] Tang J, Nie Y, Markhasin L, et al. Diffuscene: Denoising diffusion models for generative indoor scene synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 20507-20518.https://doi.org/10.48550/arXiv.2303.14207

[2] Ju X, Huang Z, Li Y, et al. Diffindscene: Diffusion-based high-quality 3d indoor scene generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 4526-4535.https://doi.org/10.48550/arXiv.2306.00519

[3] Zhang Y, Zhang H, Cheng Z, et al. SSP-IR: Semantic and Structure Priors for Diffusion-based Realistic Image Restoration[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025..https://doi.org/10.1109/TCSVT.2025.3538772.

[4] Wang C, Peng H Y, Liu Y T, et al. Diffusion models for 3D generation: A survey[J]. Computational Visual Media, 2025, 11(1): 1-28.https://doi.org/10.26599/CVM.2025.9450452.

[5] Jabri A, van Steenkiste S, Hoogeboom E, et al. DORSal: Diffusion for Object-centric Representations of Scenes et al[J]. arXiv preprint arXiv:2306.08068, 2023.https://doi.org/10.48550/arXiv.2306.08068

[6] Zhai G, Örnek E P, Wu S C, et al. Commonscenes: Generating commonsense 3d indoor scenes with

scene graph diffusion[J]. Advances in Neural Information Processing Systems, 2023, 36: 30026-30038.https://doi.org/10.48550/arXiv.2305.16283

[7] Wu Z, Li Y, Yan H, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation[J]. ACM Transactions on Graphics (ToG), 2024, 43(4): 1-17. https://doi.org/10.1145/3658188

[8] Han X, Zhao Y, You M. Scene Diffusion: Text-driven Scene Image Synthesis Conditioning on a Single 3D Model[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 7862-7870.https://doi.org/10.1145/3664647.3681678.

[9] Li X, Wu Y, Cen J, et al. iControl3D: An interactive system for controllable 3D scene generation[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 10814-10823.https://doi.org/10.1145/3664647.3680557.

[10] Gomel E, Wolf L. Diffusion-Based Attention Warping for Consistent 3D Scene Editing[J]. arXiv preprint arXiv:2412.07984, 2024.https://doi.org/10.48550/arXiv.2412.07984.

[11] Yang X, Man Y, Chen J, et al. SceneCraft: Layout-guided 3D scene generation[J]. Advances in Neural Information Processing Systems, 2024, 37: 82060-82084.https://doi.org/10.48550/arXiv.2410.09049

[12] Anciukevičius T, Xu Z, Fisher M, et al. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 12608-12618.https://doi.org/10.48550/arXiv.2211.09869.

[13] Xu Y, Chai M, Shi Z, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 4402-4412.https://doi.org/10.48550/arXiv.2212.11984

[14] Gao G, Liu W, Chen A, et al. Graphdreamer: Compositional 3d scene synthesis from scene graphs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 21295-21304.https://doi.org/10.48550/arXiv.2312.00093

[15] Gao L, Sun J M, Mo K, et al. Scenehgn: Hierarchical graph networks for 3d indoor scene generation with fine-grained geometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8902-8919.https://doi.org/10.1109/TPAMI.2023.3237577

[16] Bourigault E , Bourigault P .MVDiff: Scalable and Flexible Multi-View Diffusion for 3D Object Reconstruction from Single-View[J].IEEE, 2024.https://doi.org/10.1109/CVPRW63382.2024.00753.

[17] Yang B, Luo Y, Chen Z, et al. Law-diffusion: Complex scene generation by diffusion with layouts[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 22669-22679.https://doi.org/10.1109/ICCV51070.2023.02072

[18] Shriram J , Trevithick A , Liu L ,et al.RealmDreamer: Text-Driven 3D Scene Generation with Inpainting and Depth Diffusion[C]//2024.https://doi.org/10.48550/arXiv.2404.07199

[19] Dong Q. Surface defect detection algorithm for aluminum profiles based on deep learning[J]. Informatica, 2024, 48(13). https://doi.org/10.31449/inf.v48i13.6180

[20] Zhang G, Zhang J. High-precision photogrammetric 3d modeling technology based on multi-source data fusion and deep learning-enhanced feature learning using internet of things big data[J]. Informatica, 2025, 49(11). https://doi.org/10.31449/inf.v49i11.7137