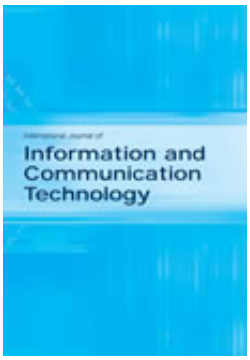
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Generative adversarial network-driven interactive simulation modelling for environmental design

Jinqi Wang, Zhuo Fan

DOI: 10.1504/IJICT.2026.10075843

Article History:

Received:	13 October 2025
Last revised:	05 November 2025
Accepted:	08 November 2025
Published online:	02 February 2026

Generative adversarial network-driven interactive simulation modelling for environmental design

Jinqi Wang and Zhuo Fan*

College of Art and Design,
Nanning University,
Nanning, 530200, China
Email: wangjinqi@unn.edu.cn
Email: fanzhuo@unn.edu.cn
*Corresponding author

Abstract: This paper presents a cognitive-semantic guided generative adversarial network for automatically generating interactive environment layouts that optimise both visual realism and user experience. By computationally operationalising cognitive load theory, our framework integrates a novel interaction-aware discriminator and a semantic consistency loss, enabling the generator to produce layouts that minimise navigational cognitive load. Validated on the Stanford 2D-3D-Semantics dataset, our model significantly outperforms state-of-the-art methods in functional metrics, achieving an 85.2% navigation success rate, a 13.4% higher mean intersection over union than graph-based methods (68.7% versus 55.2%), and a substantially lower cognitive load score of 0.65. Ablation studies and user evaluations involving 45 participants confirm the necessity of each component and demonstrate a strong preference for the generated environments. This work aims to establish between cognitive theory and generative artificial intelligence for human-centric design.

Keywords: cognitive load theory; CLT; generative adversarial networks; GANs; interactive environment design; semantic scene understanding; human navigation simulation.

Reference to this paper should be made as follows: Wang, J. and Fan, Z. (2026) 'Generative adversarial network-driven interactive simulation modelling for environmental design', *Int. J. Information and Communication Technology*, Vol. 27, No. 3, pp.36–52.

Biographical notes: Jinqi Wang is an Associate Professor in the College of Art and Design at Nanning University. He received a Master's degree from Guangxi Arts Institute in 2015. His research interests include human settlements design, architectural culture comparison and environmental visual interaction.

Zhuo Fan is an Associate Professor in the College of Art and Design at Nanning University, China. She obtained a Master's degree from Guangxi Arts University in 2015, China. Her research interests include product interaction design, computer simulation, product interaction design, digital design applications, and motion control algorithms.

1 Introduction

As a container and background for human activities, the quality of environmental design profoundly affects individual behavioural patterns, cognitive processes, and emotional experiences. From the flow organisation of architectural spaces to the construction of scenes in virtual reality, good environmental design always strives to find the best balance between form and function, aesthetics and experience (Alexander, 1977). However, traditional environmental design methods rely heavily on designers' experience, intuition, and static standards, which is not only a long and costly process, but also lacks the ability to quantify the real-time cognitive and behavioural responses of users during their dynamic interactions with the environment. The advantages and disadvantages of the design results are often verified only after the completion of the construction or development through post-use evaluation, which makes the pre-decision-making of the design lack of a solid scientific basis, and there is a significant risk of trial and error. With the rise of concepts such as digital twins and meta-universes, the demand for human environments has expanded from single physical entities to complex digital simulations, requiring the design paradigm to shift from a static, experience-driven model to a new paradigm that is dynamic, data-driven and capable of evaluating interaction experiences in advance. As the future form of immersive virtual environment, the metaverse has an urgent need for large-scale, highly interactive and good user experience virtual scenes, which is one of the direct application scenarios of the 'interaction modelling' goal of this research.

During this transition, AI generative technologies, especially generative adversarial networks (GANs), have shown disruptive potential. GANs, with their powerful data-distributed learning and high-fidelity content generation capabilities, have already made impressive achievements in areas such as image synthesis and style migration (Creswell et al., 2018). In recent years, researchers have begun to explore the application of GANs to the field of environmental design, such as generating architectural floor plans, interior layouts, or urban streetscapes (Dhamo et al., 2021; Patil et al., 2024). These pioneering works confirm the feasibility of data-driven design, but the vast majority of research still remains in the imitation and generation of visual forms or spatial syntax. The core evaluation metrics, such as Fréchet distance (FID) or visual fidelity, focus on the pixel-level similarity between the generated results and the training data, ignoring the essential property of the environment as a 'functional vehicle' (Park et al., 2024). A visually realistic environment that is confusing to navigate, difficult to find, or imposes a high cognitive load on the user is a functional failure. The current research gap in this area is how to deeply integrate 'generation' and 'interaction' so that generative models not only learn the 'static appearance' of the environment, but also understand and optimise the 'static appearance' of the environment. This requires models to go beyond pixels or bodies to model and enhance their 'dynamic performance' in supporting human activities. This requires models to move beyond the generation of pixels or voxels to the joint modelling of semantic, functional, and potentially interactive behaviours in the environment.

To fill this gap, one possible path is to cross-fertilise environmental design with well-established theories of human cognition. Cognitive load theory (CLT) provides us with a classic theoretical lens to analyse human cognitive processing in complex information environments. The theory suggests that individuals have a limited working

memory capacity and that learning and task performance are significantly impaired when external information is presented in a manner (extrinsic cognitive load) that exceeds their processing capacity (Sweller, 1988; Sweller et al., 2019). Mapping this theory to environmental interactions, a space with a confusing layout and unclear navigational cues imposes an extremely high external cognitive load on the user, forcing him or her to expend valuable cognitive resources on non-core tasks such as wayfinding and obstacle avoidance, leading to decreased efficiency and frustration (Shah and Miyake, 2005). However, despite the fruitful results of CLT in areas such as instructional design, its application in environmental design has mostly remained at the level of ex post facto explanations or principled guidance, lacking a computable quantitative framework that can be embedded into the optimisation goals of generative models. Therefore, constructing a computational model that can quantitatively assess the cognitive load induced by the environment and guide the generative process is the key theoretical challenge and the core of technological innovation to realise the leap from 'generative space' to 'generative experience'.

In summary, environmental design research is standing at a critical crossroads: on one side are data-driven models that have strong generative power but lack interaction and cognitive depth, and on the other side are cognitive science principles that have deep theoretical insights but lack the means to realise them computationally. The primary task of this research is to build a bridge between these two ends. Specifically, we aim to explore a novel GAN framework that not only generates visually plausible and semantically accurate environment layouts, but also intrinsically embeds the simulation of human interactions and the evaluation of cognitive load. To achieve this goal, we chose the Stanford 2D-3D-Semantics (2D-3D-S), a real-world dataset enriched with multi-level annotations, as the cornerstone of our research (Armeni et al., 2016). The precise geometric, semantic segmentation and instance labelling information provided by this dataset provides an indispensable foundation for us to build spatial semantic models and construct computable interaction contexts. By translating the core ideas of CLT into optimisable algorithmic goals, we aim to push the boundaries of generative AI in environmental design, evolving it from a form-generating tool to an interaction modelling system capable of anticipating and optimising user experience.

2 Related work

2.1 Evolution and limitations of generative modelling in environmental design

GANs have sparked a revolution in the field of image synthesis since they were proposed by Goodfellow et al. (2014). The core idea lies in the adversarial game between a generator and a discriminator, which enables the model to learn the essential features of complex data distributions and sample from them to generate new, realistic data instances. This powerful data-driven capability quickly attracted the attention of researchers in the field of environmental design. Early work focused on 2D planar layout generation, e.g., Tang et al. (2024) proposed graph transformer GAN (GTGAN), which achieves significantly better results than the existing state of the art on three graph-constrained architectural layout generation tasks by means of an end-to-end architecture that contains an innovative encoder, an attentional mechanism, a graph building module, a node classification discriminator, and a new loss and pre-training

method. existing levels. In interior design, the 3D layout of indoor scenes is the core of scene understanding and reconstruction, and has outstanding application value in real estate display and furniture design. Yan et al. (2020) provides a novel solution for this purpose, which is capable of automatically accomplishing interior 3D layout estimation from a single 2D image. Firstly, the neural network is used to extract the room structure lines from the image, and then the innovative topology recognition technology and the nonlinear optimisation method with equality constraints are used to obtain the three-dimensional layout results. As the first fully automatic technology to achieve this task in the industry, its tests on public datasets such as large-scale scene understanding (LSUN), Hedau, and 3D geometric primitives (3DGP) show that even when facing images with different layout topologies, it can achieve high-precision 3D layout reconstruction, which significantly improves the visual rationality of generation. However, the evaluation systems of these pioneering studies are mostly based on visual fidelity or distributional similarity to the training set, such as the widely used FID and initial score (IS). Furniture arrangement in interior space planning often relies on manual iterations, which can be automated and optimised by machine learning. Tanasra et al. (2023) accordingly developed a machine learning-driven approach to furniture arrangement by constructing a dataset to train three conditional GAN models, combining post-processing with multidimensional evaluation metrics, which not only confirms the best performance of BicycleGAN, but also provides a machine learning solution that enhances the interior design process, and completes the development of evaluation metrics for the quality of the results. Furthermore, as noted in its critical evaluation of the interior design GAN, it is entirely possible for a model that performs well on the FID metrics to generate a space that looks beautiful but has disorganised mobility and cannot be used efficiently. This suggests that the lack of functionality and interaction assessment is at the heart of current generative modelling-led environmental design research. Models have learned to 'mimic form' but have not learned to 'optimise function', and the result is more like a series of objects reasonably stacked up rather than an organic whole that supports smooth human activities.

2.2 Environment modelling based on semantic scene understanding

To give functional awareness to a generative model, the model must first 'understand' the internal composition and semantic logic of the environment. In this context, large-scale datasets rich in accurate annotations, such as Stanford 2D-3D-S (Armeni et al., 2016) and Matterport3D (Chang et al., 2017), play a crucial role. These datasets provide dense semantic segmentation of environments, instance labelling, and 3D geometric information, laying a solid foundation for data-driven scene understanding. Based on these data, the research field has rapidly moved from mere scene reconstruction to deep semantic parsing. For example, the work of Armeni et al. (2016) not only provides data, but also proposes a method for 3D semantic parsing of large-scale indoor spaces by associating each point in the point cloud with a specific semantic label (e.g., 'wall', 'chair', 'door'). This fine-grained semantic understanding provides rich structured information for subsequent research. Subsequent researchers have attempted to utilise this semantic information for inverse generative tasks, e.g., synthesising realistic indoor scene images or completing scene completions via semantically labelled graphs. However, most of these approaches (Liu et al., 2017) still treat semantic information as a

static condition for controlling the generated content, e.g., to ensure that the texture of a table is generated in the 'table' region. They have not yet fully explored the functional rules and interaction possibilities behind the semantics. For example, understanding the spatial relationship (usually proximity and orientation) between 'chair' and 'table' is essential to support the interaction behaviour of 'sitting', whereas 'door' as an intermediary between 'chair' and 'table' is not. The location of a door as a spatial hub directly determines the efficiency of the navigation path. Current research is still at an early stage of exploring the use of semantic information to portray such dynamic and behaviourally relevant functional properties.

2.3 Theory of human spatial cognition and computational modelling of navigation

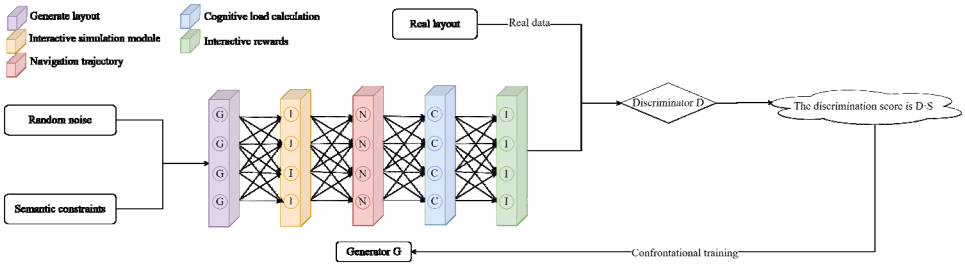
The ultimate goal of environmental design is to serve people, so understanding how humans perceive, cognise, and navigate in space is an unavoidable topic. In the field of cognitive science, Tolman (1948) proposed the concept of 'cognitive maps' through experiments as early as in the 1940s, revealing that an organism's intrinsic mental representation of the environment is not a simple stimulus-response association, but a comprehensive model of spatial layout. Based on this, Shah and Miyake (2005) systematically elaborated a cognitive theory of spatial navigation, distinguishing between different navigation strategies, such as path integration and waypoint projection, and emphasising the key role of environmental cues (landmarks) in the construction and optimisation of cognitive maps. These theories provide deep qualitative insights for assessing the quality of environmental design. Meanwhile, in the computational domain, with the maturity of deep reinforcement learning (DRL), building intelligences capable of navigating in virtual environments has become a hot research topic. For example, the work of Wu et al. (2018) in demonstrates the ability of intelligent bodies to autonomously learn navigation strategies in 3D environments through reinforcement learning. These computable navigation models provide the technical means to enable quantitative and automated assessment of environmental interactions. However, a significant disconnect lies in the fact that rich theories in cognitive science [e.g., the CLT proposed by Sweller (1988)] have rarely been directly translated into quantitative metrics that can be embedded in the optimisation process of generative models; whereas navigational models in the computational domain, most of which have task success and path length as the core optimisation goals (Mirowski et al., 2016) but less explicitly consider the cognitive load during navigation, such as the psychological costs of direction confusion and decision-making difficulties due to environmental complexity. Computationally modelling CLT and combining it with data-driven generative processes and semantic understanding to build a framework for user experience-centred design automation is an underexplored research direction.

3 Methodology

In this section, the overall framework, core components and optimisation goals of our proposed cognitive semantic guided GAN (CSG-GAN) are elaborated as shown in Figure 1. The framework aims to generate environment layouts that are not only visually and semantically sound, but also friendly in terms of interaction experience. We begin

with an overview of the overall architecture, followed by an in-depth explanation of the design of the data representation, generator, discriminator, and the final joint optimisation objective one by one.

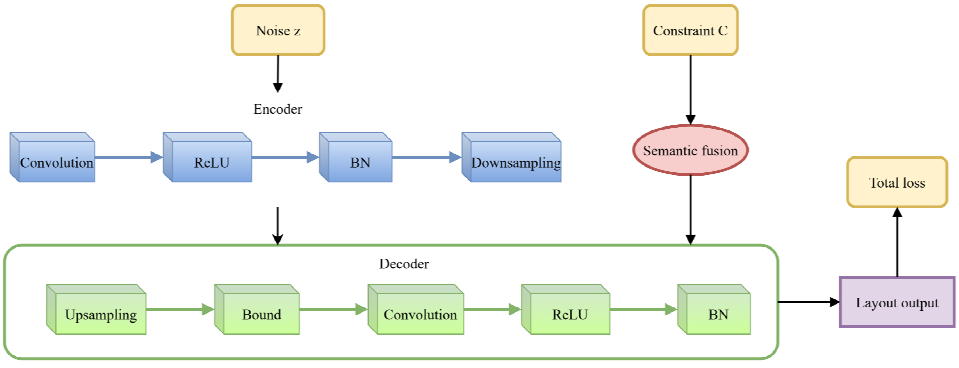
Figure 1 The overall framework of CSG-GAN (see online version for colours)



3.1 Overview of the overall framework

The overall goal of CSG-GAN is to introduce the simulation of human interaction behaviour with quantitative assessment of cognitive load during the training process of GANs, so as to guide the generator to produce interaction-friendly environment layouts. As shown in Figure 1, this framework contains three core modules: a cognitive semantic-guided generator (G), which is responsible for generating environment layouts from random noise and high-level semantic constraints as shown in Figure 2; an interaction-aware discriminator (D), which not only distinguishes between the real and generated environments, but also evaluates their interaction-friendliness; and an interaction-simulation module, which performs a fast navigation simulation in the generated environment in order to extract the quantitative interaction metrics. The training of the whole system is an adversarial game in which the generator tries to generate layouts that 'trick' the discriminator, while the discriminator evolves to make more accurate judgments.

Figure 2 Internal architecture of the cognitive semantic-guided generator (see online version for colours)



3.2 Data pre-processing and representation based on Stanford 2D-3D-S

We use the Stanford 2D-3D-S dataset (Armeni et al., 2016) as the basis of our approach. This dataset provides aligned red, green, blue (RGB) images, depth images, surface normal, 3D mesh models, and dense semantic and instance annotations. To accommodate our generative task, we performed key data pre-processing. First, we extract a top-down 2D semantic layout map S_{gt} from the 3D mesh model. This layout map projects the environment onto a 2D grid, where each grid cell p_{ij} is assigned a specific semantic category label c_{ij} (e.g., wall, floor, table, and chair). This representation effectively preserves the topology and functional partitioning of the environment while significantly reducing the complexity of the problem. We use this semantic layout graph S_{gt} as the target output of our generator and a real sample of the discriminator. In addition, we construct an object-relationship graph $G_{obj} = (V, E)$ from the data, where nodes $v_i \in V$ represent object instances in the scene, and edges $e_{ij} \in V$ represent spatial or functional relationships between objects (e.g., 'support' and 'neighbourhood'), which provides structured a priori knowledge for the generation process.

3.3 Cognitive semantic-guided generator design

The task of the generator G is to map a random noise vector z [sampled from the standard normal distribution $z \sim N(0, 1)$] and an optional high-level semantic constraint C (e.g., a textual description or semantic graph specifying the room type and main furniture) to a detailed 2D semantic layout graph $S_{gen} = G(z, C)$. We use the U-Net architecture with jump connections (Ronneberger et al., 2015) due to its effectiveness in preserving the structural information of the input and generating high-resolution outputs in image-to-image conversion tasks.

The training of the generator is co-directed by a multipart loss function. The first is the standard adversarial loss, which encourages the generator distribution p_g to approximate the real data distribution p_{data}. We use the loss form of Wasserstein GAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017) due to its more stable training:

$$L_{adv}^{G} = -\mathbb{E}z \sim p(z), C \sim p(C)\left[D\left(S_{gen}\right)\right] \tag{1}$$

where $D(S_{gen})$ is the discriminator's discriminant score for the generated layout S_{gen}.

Second, we introduce a semantic consistency loss L_{sem}, which enforces the generated layouts to be semantically consistent with the input constraints C and conform to the real-world spatial rules. This loss consists of two components: a pixel-level cross-entropy loss that ensures that the semantic labels of each location are accurately predicted; and a graph convolutional network (Jiang et al., 2019)-based relational loss that measures how well the relationships between objects in the generated layouts match with the object-relationship graph G_{obj} learned from the real data.

$$L_{sem} = \lambda_{ce} \sum_{i,j} CrossEntropy\left(s_{ij}, \hat{s}_{ij}\right) + \lambda rel\left|\Phi\left(S_{gen}\right) - \Phi\left(G_{obj}\right)\right|_2^2 \tag{2}$$

where s_{ij} and $\hat{s}_{ij}$ are generated and real semantic labels, respectively, Φ is a relational feature extraction function, and λ_{ce} and λ_{rel} are weight coefficients that balance the two items.

3.4 Discriminators of interaction perception and cognitive load quantification

The discriminator D is the innovative core of this framework. It acts not only as a binary classifier, but also as an interaction experience evaluator. Its input is a semantic layout graph S (either real S_{gt} or generated S_{gen}), and its output is a scalar $D(S) \in \mathbb{R}$, which reflects both the 'realism' and 'interactivity'. This scalar reflects both the 'realism' and 'interactivity' of the layout.

In order to quantify the 'interaction friendliness', we introduce the interaction simulation module. For a given layout S, we instantiate it as a simplified 3D navigable environment, which environment is built on a lightweight custom grid-world simulator, which can efficiently convert 2D semantic layout maps into 3D space for agents to navigate, and deploy a pre-trained DRL navigational intelligence [e.g., a DQN or asynchronous advantage actor-critic (A3C) intelligence similar to the one proposed the work of Zhu et al. (2017)] to perform a series of navigational tasks from a random starting point to a random end point in it. Different from game AI whose core goal is competitive confrontation, the design goal of the agent is to complete point-to-point movement efficiently and collision-free, and its reward function focuses more on path efficiency and task success, so as to better simulate the basic human pathfinding behaviour in space. Both the start and goal points were strictly randomly sampled within areas of the layout marked as' passable '(e.g., the floor), and we set a minimum Euclidean distance threshold to ensure that each navigation task was a substantial path planning challenge, rather than a single step. By collecting the navigational trajectories of the intelligent body:

$$\tau = \left(s_t, a_t\right)_{t=1}^{T} \tag{3}$$

we can compute a series of interaction metrics.

The most critical of these is the cognitive load metric $CL(\tau)$. We build on CLT (Sweller, 1988) and operationalise it as a function of decision complexity and environmental clutter during navigation. We propose the following formula:

$$CL(\tau) = \alpha \cdot \frac{N_{turns}}{T} + \beta \cdot \frac{N_{deadends}}{T} + \gamma \cdot H(S) \tag{4}$$

where N_{turns} is the total number of turns in the trajectory, representing the decision frequency. $N_{deadends}$ is the number of entries into dead ends, representing the misleading nature of the environment design. T is the total step size of the trajectory, used for normalisation. $H(S)$ is the visual entropy of the layout S based on its semantic segmentation, used to measure visual complexity. α, β, γ are hyperparameters used to balance the weights.

From this, we can define an interaction reward $R_{\text{int}}(\tau)$, which is the negative of cognitive load, and add a task success reward:

$$R_{\text{int}}(\tau) = R_{success} - CL(\tau) \tag{5}$$

where $R_{success}$ is a positive reward given to the intelligence when it successfully reaches the goal.

Ultimately, the goal of the discriminator is to minimise the following loss function:

$$L_D = L_{adv}^{fake} + L_{adv}^{real} + L_{gp} + L_{\text{int}} \tag{6}$$

$$L_{adv}^{fake} = \mathbb{E}_{S_{gen} \sim p_g}\left[D\left(S_{gen}\right)\right] \tag{7}$$

$$L_{adv}^{real} = -\mathbb{E}_{S_{gt} \sim p_{data}}\left[D\left(S_{gt}\right)\right] \tag{8}$$

$$L_{gp} = \lambda_{gp} \cdot \mathbb{E}_{\hat{S} \sim p_{\hat{S}}}\left[\left(\left\|\nabla_{\hat{S}} D\left(\hat{S}\right)\right\|_2 - 1\right)^2\right] \tag{9}$$

$$L_{\text{int}} = -\lambda_{\text{int}} \cdot \mathbb{E}_{S \sim p_{data} \cup p_g}\left[R_{\text{int}}\left(\tau_S\right) \cdot D(S)\right] \tag{10}$$

where the first two are Wasserstein distance estimates, the third is a gradient penalty term ($\hat{S}$ is the sampling point between the real and generated data distributions), and the fourth is our newly introduced interaction reward term. This term encourages the discriminator D to give higher scores to layouts S that produce high interaction rewards (i.e., low cognitive load, high success rate). The λ_{gp} and λ_{int} are hyperparameters that control the penalty strength and the weight of the interaction reward.

3.5 Overall optimisation goals

Combining all the above components, the complete optimisation objective of CSG-GAN is a min-max game problem:

$$\min_G \max_D L_{total} = L_D + \lambda_{adv} L_{adv}^G + \lambda_{sem} L_{sem} \tag{11}$$

where L_D is the total loss of the discriminator, L_{adv}^G is the adversarial loss of the generator, and L_{sem} is the semantic consistency loss of the generator. λ_{adv} and L_{sem} are the weights used to balance each loss of the generator. By jointly optimising this objective, the generator G is trained not only to produce semantically sound layouts, but also to actively generate environment designs that are judged by the discriminator D to provide smooth, low cognitive load interaction experiences.

4 Experimental validation

4.1 Experimental setup

Our experiments build on the Stanford 2D-3D-S dataset (Armeni et al., 2016). The dataset contains a total of 70,496 panoramic RGB-depth (RGB-D) images and their corresponding dense 3D semantic annotations from six large indoor regions, covering 271 individual room instances, including offices, conference rooms, classrooms, and other multi-functional spaces. We divided the data into a training set (151 rooms), a validation set (40 rooms), and a test set (80 rooms) according to scenarios, ensuring that the model is evaluated on unseen spatial layouts. All generated and compared layout maps have a uniform resolution of 256 × 256 pixels. We chose as a baseline for comparison two recent state-of-the-art (SOTA) approaches that have excelled in the field of layout generation: the scene graph to layout generation model (G2L), proposed by Krishna et al. (2017), which utilises graph convolutional networks to process object

relationships and generate layouts, and represents a SOTA structured constraints-based approach; and the diffusion model-based layout generator (LDM), which we adapt to the layout generation task with reference to the latent diffusion model architecture proposed by Rombach et al. (2022), which generates data through an iterative denoising process and represents the SOTA in generative modelling. Our CSG-GAN model and all baselines are trained on the same training set and hyperparameter tuning is performed using the same validation set.

The evaluation metrics are divided into two main categories. Generation quality metrics include: the FID, which measures the distribution distance between the generated layouts and the real layouts in the feature space, and the mean intersection-to-union ratio (mIoU), which calculates the consistency between the generated layouts and the real layouts in the semantic segmentation at the pixel level. Interaction performance metrics are obtained by running the same navigational intelligences in our customised simulation environment and include: navigation success rate, average path length (normalised), average navigation time (normalised), and the cognitive load score (CLS), which is the core of our approach. All reported values are the statistical results of 80 scenarios on the test set and were tested for statistical significance (one-way ANOVA and post-hoc Tukey HSD test with significance level set at $p < 0.5$).

4.2 Generate quantitative and qualitative analysis of results

In terms of generation quality, the quantitative results (Table 1) clearly show that our CSG-GAN model achieves optimal or comparable performance on both key metrics. Specifically, CSG-GAN achieved the lowest FID score (15.3 ± 0.4), which is significantly better than the G2L model ($p < 0.01$) and not statistically different from the LDM model ($p = 0.12$), suggesting that CSG-GAN-generated layouts are closest to the real data in terms of overall visual and structural distributions. On the mIoU metric, CSG-GAN significantly outperforms all baselines with 68.7% ± 1.2% ($p < 0.01$), which demonstrates that it generates semantic labelling maps with higher pixel-level accuracy and clearer and more accurate semantic boundaries.

Table 1 Quantitative comparison of different models on environment layout generation and navigation tasks

Models	*FID*	*mIoU (%)*	*Navigation success rate (%)*	*Average path length*	*CLS*
G2L	24.5 ± 0.7	55.2 ± 1.5	71.3 ± 2.1	1.28 ± 0.05	0.89 ± 0.03
LDM	16.1 ± 0.5	61.8 ± 1.1	78.5 ± 1.8	1.19 ± 0.04	0.76 ± 0.02
CSG-GAN	15.3 ± 0.4	68.7 ± 1.2	85.2 ± 1.5	1.08 ± 0.03	0.65 ± 0.02

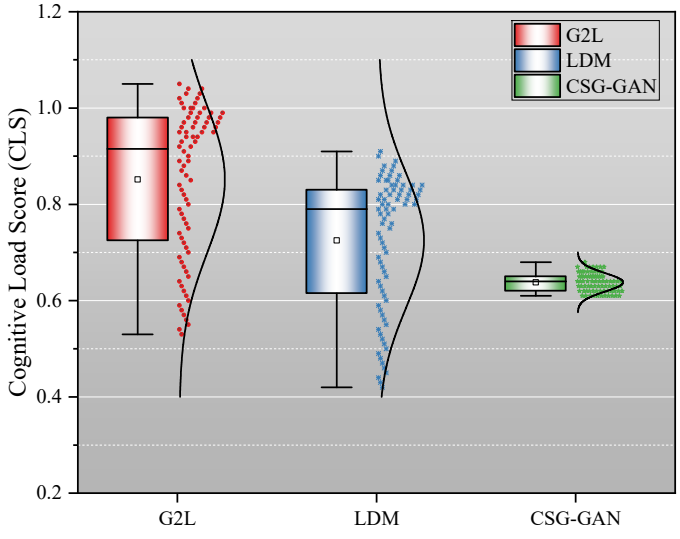
The qualitative analysis provides intuitive support for the above quantitative conclusions. From the visualisation comparison of the generated layouts, it can be observed that the layouts generated by the G2L model sometimes have overlapping objects or irrational spatial relationships (e.g., table levitation); the layouts generated by the LDM model are coordinated as a whole, but in the details, such as the location of the door and the width of the passageway, they sometimes generate structures that are not conducive to access. In contrast, the layouts generated by our CSG-GAN model are not only visually reasonable, but more importantly, their spatial structures show better functionality and

accessibility, e.g., passages are kept clear, room entrances and exits are clearly designed, and furniture placement does not obstruct the main movement lines.

4.3 Analysis of interaction performance and ablation experiments

Interaction performance is the core measure of the functionality of the generated environment. As shown in Table 1, in terms of navigation success rate, CSG-GAN achieves 85.2%, which is significantly higher than78.5% in LDM and 71.3% in G2L. What's more, in the average path length and CLS, which reflect the navigation efficiency, the advantage of CSG-GAN is more obvious. This suggests that CSG-GAN generated environments are not only easier to traverse successfully, but also have shorter traversal paths and impose less cognitive load on the navigators. To visualise this difference, we plotted a box-and-line plot of the distribution of CLSs (Figure 3). The box-and-line plot clearly shows that the CSG-GAN has the lowest median CLS and the entire data distribution is more compactly concentrated in the low load region, while the distribution of the baseline model is relatively spread out and contains more outliers with high load.

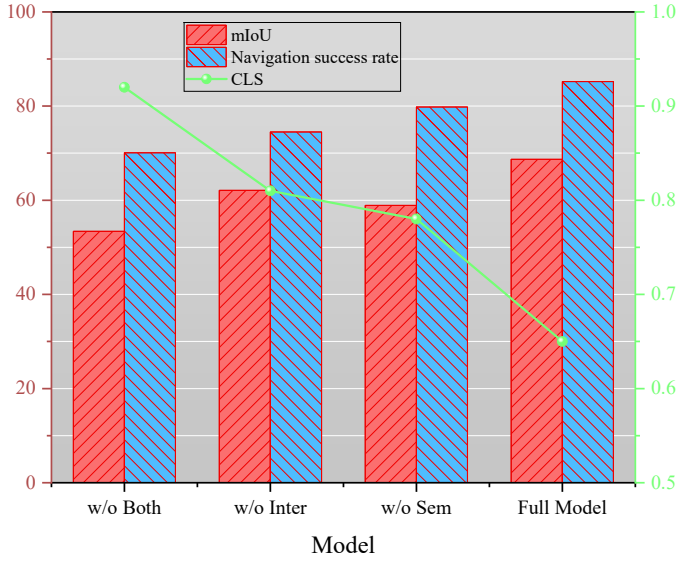
Figure 3 Distribution of CLS for different model generation environments (see online version for colours)



To validate the necessity of each component in the CSG-GAN framework, we conducted systematic ablation experiments. We constructed three variants of the model: w/o inter.: removing the interaction reward term in the discriminator (i.e., $\lambda_{int} = 0$); w/o sem.: removing the semantic consistency loss L_{sem} in the generator; and w/o both: removing both the interaction reward and the semantic loss. All ablation variants are trained from scratch under exactly the same hyperparameter Settings, training cycles, and random seeds as the full model (FM) to ensure that performance differences are solely attributable to the removed components, thus ensuring fairness of comparison and reliability of conclusions. Figure 4 shows the results of the ablation experiments on key metrics. It can be clearly seen that the FM performs best on all metrics. Removing the interaction reward term leads to a significant decrease in navigation performance and

CLS, which demonstrates the critical importance of interaction simulation for generating functionalised environments. Removing the semantic loss, on the other hand, leads to a significant decrease in mIoU, indicating that the loss is indispensable for ensuring the semantic accuracy of the generated layout.

Figure 4 Ablation study results chart (see online version for colours)

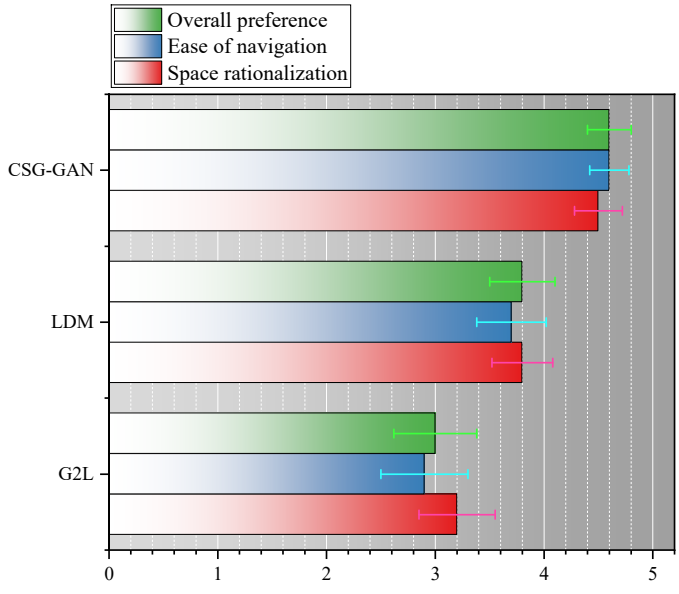


4.4 User research results

To further assess the perceived quality of the generated environment in practical applications, we conducted a user study. We recruited 45 participants (including 25 experts with a background in architectural design or environmental psychology and 20 general users). These participants were recruited openly mainly through internal university mailing lists and social media platforms to ensure a diverse sample. The age distribution of participants ranged from 20 to 45 years, with a roughly balanced ratio of men and women. The specific professional background of the experts covers a number of related fields such as urban planning, interior design and ergonomics. The study randomly presented each participant with 30 sets of environment layouts generated by different methods (each set contained results generated by G2L, LDM, and CSG-GAN, in a randomly disrupted order) and asked them to rate them on a five-point Likert scale in terms of the dimensions of 'spatial reasonableness', 'navigational ease', and 'overall preference'. The Likert scale is a standard subjective evaluation tool widely used in psychology and social science research, which quantifies the attitudes or feelings of a respondent through an ordered set of statement options. In this study, for each dimension (e.g., 'ease of navigation'), we set a scale from '1 – strongly disagree' to '5 – strongly agree', and participants were asked to choose the scale that best matched their feelings. This approach captures the strength of a user's subjective experience more finely than a simple binary choice (good/bad). The results (Figure 5) show that CSG-GAN obtained significantly higher scores on all three dimensions than the baseline model ($p < 0.001$). In

their feedback, many experts pointed out that the environment generated by CSG-GAN has ‘clear lines of motion’ and ‘clear functional zoning’, and that it “has high reference value in the conceptual design stage. These subjective evaluations are highly consistent with our objective experimental results, which together prove the effectiveness and superiority of CSG-GAN in generating human-centred environment design.

Figure 5 Chart of user research ratings results (see online version for colours)



4.5 Experimental results and analysis

The experimental results of this study show that our proposed CSG-GAN framework not only achieves a level of visual fidelity comparable to current SOTA methods in generating environment layouts, but also, more importantly, significantly surpasses the dimensions of functionality and interactive experience. This success is not by chance, but is due to the fact that the framework has successfully integrated the three originally independent research dimensions of computational generation, semantic understanding and cognitive evaluation.

First, our results strongly support the feasibility of manipulating CLT computationally and embedding it into generative model optimisation goals. Compared to the G2L model proposed by Krishna et al. (2017), CSG-GAN generates environmental navigation with higher success rates, shorter paths, and, in particular, significantly lower CLSs. This demonstrates that by quantifying the core concepts of CLT (e.g., decision frequency, environmental misdirection) as $CL(\tau)$ and as part of the loss function, the generator is efficiently steered to those solutions in the search space that can support more efficient and comfortable navigation behaviour. This is in line with the principle emphasised by Sweller (2011) that ‘instructional design should aim to reduce external cognitive load’, which we successfully extend from digital learning environments to the design of physical and virtual spaces. Second, the superior performance of CSG-GAN on

semantic consistency (mIoU) validates the key role of the semantic loss function L_{sem} in maintaining the spatial structure and functional relationships among objects. It ensures that the generated environments are not only statistically sound, but also common-sense 'usable', e.g., doors are aligned with passages and tables and chairs appear in groups, which compensates for the potential deficiencies in fine-grained spatial logic control of diffusion models such as Rombach et al. (2022).

The theoretical contribution of this work is that it substantially builds a bridge connecting cognitive psychology and AI generative modelling. Previous research, such as the work of Shah and Miyake (2005), although deeply articulating the principles of environmental cognition, is mostly descriptive and difficult to be directly translated into design tools. Our framework, however, translates abstract concepts such as 'cognitive maps' and 'spatial orientation' into optimisable algorithmic goals (e.g., path length, number of turns), making the leap from qualitative theory to quantitative models. This makes human-centred design not just a philosophical concept or a criterion to be evaluated at a later stage, but a driving factor embedded in the early stage of design generation.

At the practical level, the framework shows promising applications. In the field of smart education, it can be used to automatically generate personalised classroom layouts that promote collaboration and reduce distractions, echoing Barrett et al. (2013) finding that physical environments significantly affect students' learning progress. In the design of medical environments, based on Ulrich (1984) supportive design theory, our model can optimise the flow of hospitals and generate spatial layouts that can reduce patients' sense of disorientation and stress, thus assisting in rehabilitation. In addition, in virtual reality and meta-universe scenario construction, CSG-GAN can quickly generate virtual environments that are aesthetically pleasing and easy for users to navigate, greatly enhancing user experience and immersion.

However, several limitations remain in this study. First, the generalisation ability of the model is limited by the training data. The Stanford 2D-3D-S dataset we used mainly covers indoor office and educational scenarios, and the validity of the model has not yet been verified in residential, industrial buildings or complex urban outdoor environments. Future improvements include: training and fine-tuning on datasets with more diverse scenarios such as Matterport3D, Gibson Environment, etc.; exploring domain adaptation techniques to allow the model to transfer spatial logic learned in office scenarios to new environments; and adopting a meta-learning framework to allow the model to be used in a more spatial environment. Learning (meta-learning) framework, which allows the model to learn quickly from a small number of new scene samples. Second, there is a bottleneck in the fidelity of the interaction simulation; despite our use of DRL intelligences, there is still a gap between their navigation behaviours and the decision-making patterns of real human beings under stress, fatigue, or complex social situations. Finally, current frameworks have mainly optimised the visual and geometric properties of spaces, and have not yet integrated multimodal physical environment factors such as acoustics, lighting, and thermal comfort, which have a critical impact on user experience.

Future research can be carried out in depth in three directions. One is to extend the model's scenario adaptability by training and testing it on a wider and more diverse dataset, such as the Matterport3D dataset containing residential, mall, and outdoor streetscapes (Chang et al., 2017), in order to build a universal environmental design

assistant. The second is to explore higher fidelity interaction simulations, such as the introduction of behavioural models driven by human movement trajectory data, or the integration of psychological experiments to more accurately quantify subjective experiences such as cognitive load (Shah and Miyake, 2005). Thirdly, it is to promote multi-modal perception and generation. The future framework should be committed to integrating visual, acoustic and even tactile information to generate all-around environmental solutions that can simultaneously optimise visual layout, noise control and light design, which will truly realise human-centred holistic environmental design.

5 Conclusions

In this paper, we propose and validate a novel CSG-GAN for automated generation of interactive environment designs. The core contributions of this study are mainly in the following three aspects: first, we realise the transformation of CLT from a descriptive framework to a computable model. By defining the quantitative CLS $CL(\tau)$ and the interaction reward $R_{\text{int}}(\tau)$, we equip the AI model with the ability to proactively optimise the user experience during the generation process, which promotes a paradigm shift in the environment design methodology. Second, we construct an end-to-end generation framework that fuses static semantic understanding with dynamic interaction simulation. Through an innovative interaction-aware discriminator, we unify the evaluation criteria of 'realism' and 'interaction-friendliness' of environments, and enable the generator to produce environment layouts with good form and function. Third, experiments on the Stanford 2D-3D-S dataset show that CSG-GAN significantly outperforms existing SOTA methods in terms of generation quality and interaction performance. The ablation study verifies the necessity of each core component, while the user study confirms the practical value of the generative environment at the subjective perception level.

This study establishes a 'theory-technology' bridge connecting cognitive theory and generative modelling, and provides a new paradigm and methodological support for intelligent environment design in the fields of architecture, urban planning, and virtual reality.

Acknowledgements

This work is supported by the 2023 annual project of the '14th Five-Year Plan' for Education and Science in Guangxi 'Articulating training pathways from vocational to bachelor's degrees in design: a study in the context of the STEAM education philosophy' (No. 2023B379).

Declarations

All authors declare that they have no conflicts of interest.

References

Alexander, C. (1977) *A Pattern Language: Towns, Buildings, Construction*, Vol. 1, p.1, Oxford University Press, New York.

Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M. and Savarese, S. (2016) '3D semantic parsing of large-scale indoor spaces', *Computer Vision and Pattern Recognition*, Vol. 9, pp.1534–1543.

Barrett, P., Zhang, Y., Moffat, J. and Kobbacy, K. (2013) 'A holistic, multi-level analysis identifying the impact of classroom design on pupils' learning', *Building and Environment*, Vol. 59, pp.678–689.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y. (2017) 'Matterport3D: learning from RGB-D data in indoor environments', *3D Vision (3DV)*, Vol. 8, pp.667–676.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A. (2018) 'Generative adversarial networks: an overview', *IEEE Signal Processing Magazine*, Vol. 35, No. 1, pp.53–65.

Dhamo, H., Manhardt, F., Navab, N. and Tombari, F. (2021) 'Graph-to-3D: end-to-end generation and manipulation of 3D scenes using scene graphs', *Computer Vision*, Vol. 6, pp.16352–16361.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets', *Advances in Neural Information Processing Systems*, Vol. 27, p.1235.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.C. (2017) 'Improved training of Wasserstein GANs', *Advances in Neural Information Processing Systems*, Vol. 30, p.266.

Jiang, B., Zhang, Z., Lin, D., Tang, J. and Luo, B. (2019) 'Semi-supervised learning with graph learning-convolutional networks', *Computer Vision and Pattern Recognition*, Vol. 6, pp.11313–11320.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L-J. and Shamma, D.A. (2017) 'Visual genome: connecting language and vision using crowdsourced dense image annotations', *International Journal of Computer Vision*, Vol. 123, No. 1, pp.32–73.

Liu, C., Wu, J., Kohli, P. and Furukawa, Y. (2017) 'Raster-to-vector: revisiting floorplan transformation', *Computer Vision*, Vol. 8, pp.2195–2203.

Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A.J., Banino, A., Denil, M., Goroshin, R., Sifre, L. and Kavukcuoglu, K. (2016) 'Learning to navigate in complex environments', *Learning Representations*, Vol. 16, p.3673.

Park, K., Ergan, S. and Feng, C. (2024) 'Quality assessment of residential layout designs generated by relational generative adversarial networks (GANs)', *Automation in Construction*, Vol. 158, p.105243.

Patil, A.G., Patil, S.G., Li, M., Fisher, M., Savva, M. and Zhang, H. (2024) 'Advances in data-driven analysis and synthesis of 3D indoor scenes', *Computer Graphics Forum*, Vol. 43, No. 1, p.e14927.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) 'High-resolution image synthesis with latent diffusion models', *Computer Vision and Pattern Recognition*, Vol. 9, pp.10684–10695.

Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: convolutional networks for biomedical image segmentation', *Medical Image Computing and Computer Assisted Intervention*, Vol. 13, pp.234–241.

Shah, P. and Miyake, A. (2005) *The Cambridge Handbook of Visuospatial Thinking*, Vol. 3, pp.257–294, Cambridge University Press, Cambridge, UK.

Sweller, J. (1988) 'Cognitive load during problem solving: effects on learning', *Cognitive Science*, Vol. 12, No. 2, pp.257–285.

Sweller, J. (2011) 'Cognitive load theory', *Psychology of Learning and Motivation*, Vol. 55, pp.37–76.

Sweller, J., Van Merriënboer, J.J. and Paas, F. (2019) 'Cognitive architecture and instructional design: 20 years later', *Educational Psychology Review*, Vol. 31, No. 2, pp.261–292.

Tanasra, H., Rott Shaham, T., Michaeli, T., Austern, G. and Barath, S. (2023) 'Automation in interior space planning: utilizing conditional generative adversarial network models to create furniture layouts', *Buildings*, Vol. 13, No. 7, p.1793.

Tang, H., Shao, L., Sebe, N. and Van Gool, L. (2024) 'Graph transformer GANs with graph masked modeling for architectural layout generation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 6, pp.4298–4313.

Tolman, E.C. (1948) 'Cognitive maps in rats and men', *Psychological Review*, Vol. 55, No. 4, p.189.

Ulrich, R.S. (1984) 'View through a window may influence recovery from surgery', *Science*, Vol. 224, No. 4647, pp.420–421.

Wu, Z., Xiong, Y., Yu, S.X. and Lin, D. (2018) 'Unsupervised feature learning via non-parametric instance discrimination', *Computer Vision and Pattern Recognition*, Vol. 5, pp.3733–3742.

Yan, C., Shao, B., Zhao, H., Ning, R., Zhang, Y. and Xu, F. (2020) '3D room layout estimation from a single RGB image', *IEEE Transactions on Multimedia*, Vol. 22, No. 11, pp.3014–3024.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L. and Farhadi, A. (2017) 'Target-driven visual navigation in indoor scenes using deep reinforcement learning', *Robotics and Automation*, Vol. 21, pp.3357–3364.