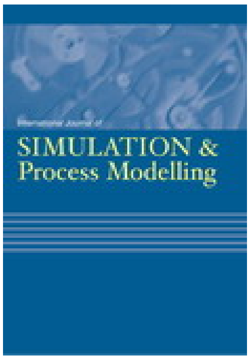
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123
https://www.inderscience.com/ijspm

Multimodal transformer-driven consistent environment design generation simulation modelling

Zhuo Fan, Jinqi Wang

DOI: 10.1504/IJSPM.2026.10075345

Article History:

Received:	14 October 2025
Last revised:	04 November 2025
Accepted:	19 November 2025
Published online:	06 March 2026

Multimodal transformer-driven consistent environment design generation simulation modelling

Zhuo Fan and Jinqi Wang*

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

Abstract: Automated generation of physically plausible 3D environments is a key challenge in digital twins, the metaverse, and robot simulation. Current methods focus mainly on visual fidelity, often overlooking functional and physical rationality, limiting direct applicability to simulation tasks. To address this, we propose a multimodal transformer-based framework for environment design generation and consistency simulation. Utilising a cross-modal attention mechanism, our model integrates textual descriptions with prior knowledge from real 3D scenes. It incorporates fine-grained physical constraint losses – including collision avoidance, support relations, and spatial accessibility optimisation – during training to explicitly model physical consistency. Experiments on the Matterport3D dataset show our method outperforms existing baselines in visual quality and layout rationality. Notably, it shows significant gains in physical consistency: collision volume is greatly reduced, and navigation success reaches 89%, affirming high simulability and practicality of the generated environments.

Keywords: multimodal transformer; environment design generation; physical consistency; simulation modelling; microphysically constrainable.

Reference to this paper should be made as follows: Fan, Z. and Wang, J. (2026) ‘Multimodal transformer-driven consistent environment design generation simulation modelling’, *Int. J. Simulation and Process Modelling*, Vol. 23, No. 1, pp.34–43.

Biographical notes: Zhuo Fan is a Lecturer 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.

Jinqi Wang is an Associate Professor in the College of Art and Design at Nanning University, China. He obtained a Master’s degree from Guangxi Arts University in 2015, China. His research interests include human settlement environment design, architectural culture comparison and environmental visual interaction.

1 Introduction

With the rapid development of digital twins, metaverse, autonomous driving simulation and embodied intelligence, the demand for high-fidelity, large-scale, and interactive virtual environments is growing at an unprecedented rate. Traditional manual environment modelling methods have become a key bottleneck in restricting the iteration and application of related technologies due to their time-consuming and costly nature (Alatan et al., 2007). Therefore, the development of intelligent methods that can automatically generate realistic, reasonable and fully functional virtual environments has become a core challenge and urgent need at the intersection of computer graphics, computer vision and artificial intelligence. An ideal environment generation system not only needs to create visually appealing scenarios, but also needs to ensure its intrinsic physical rationality and functional consistency. For example, indoor environments generated for robotics training must ensure collision-free navigation of intelligences (Savva et al., 2019); design solutions generated for architectural previews must have ergonomic and structurally stable furniture layouts. This ‘consistency’ requirement, which goes beyond apparent visual quality to the functional properties of the environment, elevates the digital environment from a purely ornamental object to a computable, interactive, and dependable simulation substrate that can reliably support downstream tasks.

In recent years, data-driven generative models, especially generative adversarial networks (Goodfellow et al., 2014) and variational self-encoders (Lopez et al., 2018), have achieved notable successes in the generation of images, videos and even 3D shapes. Researchers have also attempted to migrate these techniques to 3D scene generation tasks. Much of the early work focused on learning the geometric layout of a scene from 2D images or 3D scanned data with coexisting relationships between objects, and thus generating new scene layouts (Ritchie et al., 2019). However, these methods rely heavily on statistical regularities in large-scale labelled data, and their optimisation goals usually focus on improving the visual fidelity of the generated results or the distributional similarity to the training data. A fundamental limitation is that they lack explicit modelling of real-world physical laws and functional constraints. This results in generated environments that are often characterised by physical anomalies such as objects penetrating each other, furniture hovering in the air, and blocked pathways. Such 'inconsistencies' make the generated environments, although they may look plausible in static renderings, impossible to use in any serious simulation tasks, greatly limiting their practical value.

At the same time, revolutionary advances in the fields of natural language processing and computer vision, in particular the rise of the Transformer architecture (Vaswani et al., 2017) and its successful application in multimodal learning, have provided new paradigms for scene understanding and generation. Models such as contrastive language-image pre-training (CLIP) (Radford et al., 2021) demonstrated the power of aligning visual and linguistic concepts in a unified embedding space. Researchers have begun to explore the use of textual descriptions as guiding conditions to generate semantically compliant 3D scenes (Jain et al., 2022; Sanghi et al., 2022). These approaches are an important step towards 'controlled generation'. However, most of the current research on multimodal transformer-based scene generation still focuses on semantic relevance and visual plausibility. They are able to ensure that the generated scenes contain the objects mentioned in the textual descriptions (e.g., 'a bed' and 'a closet'), but not the precise physical properties (e.g., mass, friction) of these objects and their complex physical interactions (e.g., support, collision, stability) in the three-dimensional space. (However, there is still a lack of in-depth modelling and constraints on the precise physical properties (e.g., mass, friction) of these objects in 3D space and the complex physical interactions between them (e.g., support, collision, stability) (Paschalidou et al., 2021). In other words, the current technology is able to 'say what it means', but it has not yet reached the state of 'getting what it means' and 'making the best use of what it means'.

To overcome the above limitations, this paper explores a new research path: to incorporate physical consistency as a core design principle into a multimodal transformer-based environment generation framework. We argue that a truly intelligent environment generation system must intrinsically understand and comply with the fundamental rules of the physical world. To this end, we propose a novel generative paradigm that not only learns the visual-semantic mapping of a scene, but more importantly learns spatial common sense and physical constraints from real-world data enriched with physical information. Large-scale indoor datasets such as Matterport3D (Chang et al., 2017) provide an ideal platform for this purpose, which not only provides rich RGB-D images and accurate 3D mesh models, but also contains detailed object instance segmentation and semantic annotations, allowing us to extract precise spatial relationships and potential physical interactions between objects.

The core idea of this research is to construct a multimodal Transformer-driven generative architecture that can synergistically process and deeply fuse textual descriptions, visual cues, and geometrical and physical a priori extracted from real 3D scenes. We hypothesise that by allowing the model to explicitly engage and reason about physical relationships between objects during training (e.g., a cup should be supported by a tabletop, a chair should not be embedded in a wall), the model is able to build an implicit understanding of physical consistency into its internal representations. However, there are several key challenges to achieving this goal: first, how to design effective model structures that translate discrete, symbolic physical constraints (e.g., 'no collision', 'need to be supported') into microscopic loss functions (Battaglia et al., 2016) for easy integration into the end-to-end training process of deep neural networks. Second, how to seamlessly integrate the generated environments into a physical simulation engine (Coumans and Bai, 2016), and automate and quantitatively evaluate their physical consistency, thus forming a closed-loop 'generation-simulation-verification' feedback mechanism. Finally, how to ensure that the diversity and visual quality of the generated environments are not compromised by the introduction of physical constraints, i.e., to strike a balance between 'rationality' and 'creativity'.

By tackling these challenges, we aim to promote a paradigm shift from 'visual synthesis' to 'functional creation' in environment generation technology, and lay a solid foundation for building the next generation of simulatable and interactive intelligent digital environments.

2 Related work

2.1 Evolution and limitations of 3D scene generation

The ultimate goal of 3D scene generation is to create virtual environments that are both visually realistic and structurally sound. Early work relied on procedural generation or exemplar-based synthesis, which are highly controllable but require a lot of human intervention and are difficult to adapt to diverse needs. With the rise of deep learning, data-driven generative models have become mainstream. Generative adversarial networks proposed by Goodfellow et al. (2014) and variational self-encoders proposed by Kingma and Welling (2013) are widely used to learn the latent distribution of a scene from 2D images or 3D data. For example, some studies have been able to generate new scene layouts by learning co-occurrence probabilities and spatial layout statistics among objects from a large amount of indoor scene data (Li et al., 2019). However, these methods are largely limited by the nature of their probabilistic models, and their optimisation goals usually focus on the reconstruction accuracy at the pixel-level or point-cloud level, as well as the fit to the training data distribution. A common pitfall is that they lack explicit modelling of the functionality and physical plausibility of the scene. This leads to the generation of results that often defy physical common sense, such as interpenetration of objects, levitation of furniture, and blockage of passages. While these 'illogical' scenes may appear realistic in static snapshots, their inherent inconsistencies are exposed when placed in interactive or simulation environments, greatly limiting their utility in applications requiring full functionality, such as robot simulation and virtual reality.

2.2 Multimodal learning and the converging waves of transformer

Breakthroughs in the fields of natural language processing and computer vision, especially the transformer architecture proposed by Vaswani et al. (2017), have revolutionised multimodal comprehension and generation tasks. Transformer's self-attention mechanism is able to efficiently capture long-distance dependencies, which is well suited for dealing with complex interactions between objects in a scene. The multimodal pretraining model represented by CLIP proposed by Radford et al. (2021) successfully aligns visual and linguistic concepts into a unified semantic space, laying a solid foundation for realising text-based cross-modal generation. Researchers have rapidly applied these advances to the field of scene generation. A series of works explored how textual descriptions can be utilised as conditions for generating or editing 3D scene layouts through the transformer architecture (Sanghi et al., 2022). These methods are able to generate scenes containing relevant objects based on high-level semantic inputs such as 'a modern living room', making the leap from 'what is' to 'what is there'. The leap from 'what is' to 'what is there' has been achieved. However, most of the current research still focuses on the semantic relevance of multimodal fusion, i.e., ensuring that the content of the generated scene matches the textual description. Existing models are still insufficient in exploiting the inherent geometric accuracy and physical laws of the scene that go beyond semantics. A model may know that a sofa should coexist with a coffee table, but it may not be able to accurately calculate the distance between them in order to avoid collisions or ensure that the coffee table is stabilised by the floor. The 'floor' is stably supported (Paschalidou et al., 2021). How to allow models to deeply internalise the physical rules of 3D space while understanding the semantics is a key challenge for current research.

2.3 Physics-inspired machine learning methods

In order for machine learning models to 'understand' the physical world, an important research direction is to introduce physical knowledge or constraints into the design and training process of the models. This type of work is often referred to as physics-inspired machine learning. There are various paths to achieve this: one is to introduce physical constraint terms into the loss function, e.g., to penalise collisions by calculating the intersection ratio between the bounding boxes of objects, or to encourage stability by detecting the contact between the bottom of an object and its support surface (Zheng et al., 2020). Another, more sophisticated approach is the introduction of differentiable physics simulators that allow the forward simulation process of the physical dynamics to provide gradients to the neural network, thus guiding the model to generate results that conform to the physical laws (de Avila Belbute-Peres et al., 2018). In addition, graph neural networks are widely used to explicitly model interactions between objects, representing the scene as a graph where nodes are objects and edges are relationships between them (e.g., support, adjacency), and thus learn or reason about physical dynamics (Battaglia et al., 2016; Janner et al., 2018). Although these methods have been remarkably effective in tasks such as physics prediction and rigid-body dynamics simulation, they have mostly been applied to analytical tasks, i.e., assessing the physical plausibility or predicting the dynamics of a given scenario. How to apply these physical constraints creatively and prospectively to generative tasks, i.e., to circumvent physical irrationality at the early stage of scene construction from scratch, remains an open and challenging topic.

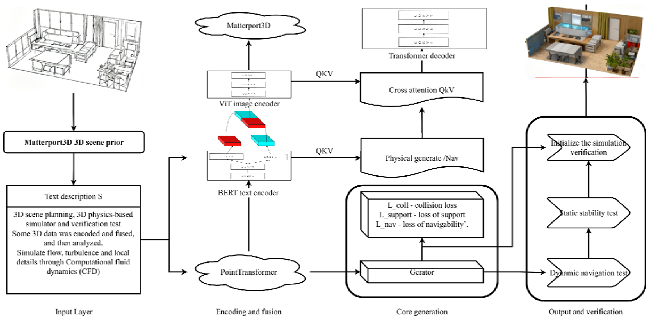
2.4 The role of indoor scene datasets and the authority of Matterport3D

Large-scale, high-quality datasets are the cornerstone for advancing all of the above research directions. In the field of indoor scene understanding and generation, several landmark datasets have been proposed one after another. For example, SUN red, green, blue – depth (RGB-D) proposed by Song et al. (2015) provides rich RGB-D images with annotations, and ScanNet proposed by Dai et al. (2017) provides dense 3D reconstruction and semantic annotations through large-scale scanning. Among them, the Matterport3D dataset proposed by Chang et al. (2017) occupies an authoritative position due to its grand scale, detailed annotation and diverse scenes. It contains 90 complete building scans covering a wide range of interior types such as residential, hotel, office, etc. It not only provides high-resolution RGB images and dense 3D mesh models, but also contains accurate camera bit positions and instance-level semantic segmentation. This multimodal and high-precision feature makes Matterport3D an ideal platform for learning scene layout a priori, spatial relationships among objects, and potential physical constraints. Compared to other datasets, the panoramic scanning feature of Matterport3D provides more complete and coherent spatial contextual information, which is crucial for generating an overall consistent environment. Matterport3D was chosen as the core dataset in this study precisely to capitalise on its all-encompassing advantages to generate 3D scenes with both semantic correctness and physical consistency by learning from real, physically present environments.

3 Methodology

In this paper, we propose a generative framework called physically constrained multimodal transformer, which aims to generate indoor 3D scenes with a high degree of physical consistency from textual descriptions. As shown in Figure 1, the framework is mainly composed of four core modules: a multimodal encoder, a cross-modal fusion Transformer, a physically constrained generator, and a consistency simulation verification module. The whole process is trained in an end-to-end manner, and the model is guided to generate environments that are not only semantically correct, but also structurally sound and simulation-ready, by explicitly embedding the rules of the physical world into the loss function. The so-called ‘end-to-end’ in this framework starts from the most original and unprocessed natural language description string, as well as (optionally) the layout sketch image provided by the user. Its ‘terminal’ is a 3D scene layout data that can be directly imported into a physical simulation engine (such as PyBullet) for consistency verification and contains complete object attributes (category, size, position, orientation). Throughout the entire process, no manual intermediate steps or post-processing optimisations are required. The model can complete the mapping from semantic understanding to a physically reasonable layout through a single forward propagation.

Figure 1 Schematic diagram of the overall framework of physically constrained multi-modal transformer (PCM-transformer) (see online version for colours)


Matterport3D
Transformer decoder
QKV
ViT image encoder
Cross attention QkV
Matterport3D 3D scene prior
QKV
Physical generate /Nav
Initialize the simulation verification
BERT text encoder
Text description S
3D scene planning, 3D physics-based simulator and verification test
Some 3D data was encoded and fused, and then analyzed.
Simulate flow, turbulence and local details through Computational fluid dynamics (CFD)
L_coll - collision loss
L_support - loss of support
L_nav - loss of navigability'.
Static stability test
PointTransformer
Gerator
Dynamic navigation test
Input Layer
Encoding and fusion
Core generation
Output and verification

3.1 *Overall framework and problem definition*

The central task of this study is to learn a mapping function that converts a given natural language description S into a 3D scene layout $\mathcal{L}$. The layout $\mathcal{L}$ can be represented as a collection of objects: $\mathcal{L} = \{o_i\}_{i=1}^{N}$, where each object $o_i = (c_i, d_i, t_i, r_i)$ contains its semantic category c_i, 3D dimension $d_i \in \mathbb{R}^3 ci$, 3D translation position $t_i \in \mathbb{R}^3$ and orientation $r_i \in \mathbb{R}^4$ (in quaternions). In order to learn real-world layout a priori and physical common sense, we utilise the Matterport3D dataset $\mathcal{D} = \{(S_j, \mathcal{L}_j)\}$ as the supervisory information, where each real-world scene $\mathcal{L}_j$ has been pre-processed to extract the objects and their attributes it contains.

Our model aims to generate a new layout $\hat{\mathcal{L}}$ that semantically matches the description S and is physically plausible by synergistically processing the textual description and the 3D scene prior learned from $\mathcal{D}$.

3.2 *Multimodal feature extraction*

- Text feature encoding: for the input natural language description S (e.g., ‘a modern living room with an L-shaped sofa and a glass coffee table’), we use a pre-trained BERT model as a text encoder. First tokenise S as a sequence, and then get the contextual embedding of each token by the BERT model. We take the final hidden state of the sequence start token `[CLS]` as the global representation of the whole sentence:

$$\mathbf{T} = \text{BERT}_{[\text{CLS}]}(S) \in \mathbb{R}^{d_t} \tag{1}$$

  where $\mathbf{T}$ is a text feature vector of dimension d_t that encodes the semantic information of the input.

- 3D scene a priori coding: in order to make our generated scenes conform to real-world spatial and physical laws, we introduce a 3D scene a priori encoder. This encoder uses a point transformer network pre-trained on the Matterport3D dataset. For a real scene $\mathcal{L}_j$, we transform it into a point cloud $\mathbf{X} = \{p_1, p_2, ..., p_M\} \in \mathbb{R}^{M \times (3+F)}$, where each point p_i contains its 3D coordinates and (optionally) semantic features F. Point transformer aggregates global contextual information through a self-attentive mechanism:

$$\mathbf{P} = \text{Point Transformer}(\mathbf{X}) \in \mathbb{R}^{d_p} \tag{2}$$

  where $\mathbf{P}$ is a global scene feature of dimension d_p that encapsulates the geometric structure of the scene, spatial relationships between objects, and potential physical constraints.

- Figure feature encoding: if the input contains layout sketches I, we use vision transformer (ViT) as an image encoder. Split the sketch I into patches and input it to ViT, also take the output of `[CLS]` markers as sketch features:

$$\mathbf{V} = \text{ViT}_{[\text{CLS}]}(I) \in \mathbb{R}^{d_v} \tag{3}$$

  where $\mathbf{V}$ is a visual feature vector of dimension d_v.

3.3 Cross-modal transformer fusion

Among the many multimodal pre-trained models, we choose the CLIP paradigm as the foundation, mainly based on its two core advantages. Firstly, CLIP, trained on a vast number of internet image-text pairs, has an extremely strong generalisation ability and zero-shot transfer potential in its visual-language embedding space. This provides a solid foundation for us to align the limited 3D data (Matterport3D) with the infinite text semantic space, enabling the model to understand and respond to text descriptions that have not appeared in the training set. Secondly, what CLIP achieves is a coarse-grained, semantic-level alignment, which is highly consistent with the goal of our task to associate text concepts (such as ‘sofa’) with three-dimensional object entities. Compared with those models that focus on fine-grained region-word alignment, the global feature representation of CLIP can better capture the overall semantics and layout intentions at the scene level, thereby more effectively guiding the generation of 3D scenes.

In order to deeply fuse information from different modalities, we designed a transformer-based cross-modal fusion module. The core of this module is to utilise the cross-attention mechanism to allow queries from one modality to retrieve relevant information from key-value pairs of another modality.

We perform cross-attention computation with textual features **T** as query and 3D scene a priori features **P** as keys and values. First, the features of different modalities are projected to a uniform dimension d_{model} by linear transformation:

$$\mathbf{Q}_t = \mathbf{T}\mathbf{W}^Q, \quad \mathbf{K}_p = \mathbf{P}\mathbf{W}^K, \quad \mathbf{V}_p = \mathbf{P}\mathbf{W}^V \tag{4}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_{model}}$ is the learnable weight matrix. Next, the attention weights and outputs are computed:

$$\text{Attention}\left(\mathbf{Q}_t, \mathbf{K}_p, \mathbf{V}_p\right) = \text{softmax}\left(\frac{\mathbf{Q}_t\mathbf{K}_p^T}{\sqrt{d_k}}\right)\mathbf{V}_p \tag{5}$$

where d_k is the dimension of the key vector. This process allows the textual description to ‘focus’ on the structure of the 3D scene relevant to its semantics. Eventually, the fused features $\mathbf{F}_{fusion}$ can be obtained through a feed-forward network (FFN) and residual concatenation:

$$\mathbf{F}_{fusion} = \text{FFN}(\text{LayerNorm}(\mathbf{T}' + \mathbf{T})) \tag{6}$$

where $\mathbf{T}'$ is the output of the cross-attention. If sketch input exists, one can similarly construct another way to cross attention and fuse its output with $\mathbf{F}_{fusion}$.

3.4 Physically constrained 3D layout generation

The fused features $\mathbf{F}_{fusion}$ are fed into a transformer-based autoregressive decoder for generating object sequences $\hat{o}_1, \hat{o}_2, ..., \hat{o}_N$. The decoder, at each step i, predicts various types of attributes (c_i, d_i, t_i, r_i) for the current object o_i based on the generated object $o_{<1}$ and the fusion features $\mathbf{F}_{fusion}$.

To ensure the physical consistency of the generated scene, we introduce several physical constraint losses in addition to the standard negative log-likelihood loss L_{recon}:

- Collision loss: penalises unreasonable penetration between objects. We approximate this using the intersection ratio between object bounding boxes:

$$L_{coll} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} \max\left(0, \delta - \text{IoU}\left(b_i, b_j\right)\right) \tag{7}$$

where b_i and b_j are the 3D bounding boxes of objects i and j respectively, IoU(·) computes their intersection ratio, and δ is a safe distance threshold that generates a loss when intersection over union (IoU) is greater than δ.

- Loss of support relationship: ensure that the object to be supported (e.g., table, chair) has sufficient contact with the support surface (e.g., floor). We calculate the ratio of the contact area between the bottom of the object and the floor:

$$L_{support} = \sum_{i \in \mathcal{O}_{sup}} \left[1 - \frac{\text{ContactArea}\left(b_i, \text{Floor}\right)}{\text{Area}\left(b_i^{\text{bottom}}\right)}\right] \tag{8}$$

where $\mathcal{O}_{sup}$ is the set of all objects to be supported, contact area calculates the contact area, and $\text{Area}\left(b_i^{\text{bottom}}\right)$ is the area of the bottom of the object.

- Loss of accessibility: to generate open, navigable space. We define it by calculating the percentage of free space area in the top view of the scene:

$$L_{nav} = -\frac{A_{free}}{A_{total}} \tag{9}$$

where A_{free} is the free space area and A_{total} is the total scene area.

The total loss function of the model is the weighted sum of each of the above losses:

$$L_{total} = L_{recon} + \lambda_{coll}L_{coll} + \lambda_{support}L_{support} + \lambda_{nav}L_{nav} \tag{10}$$

where λ_{coll}, $\lambda_{suppotr}$, λ_{nav} are hyperparameters to balance the importance of each loss. The hyperparameters in the loss function, including the safety distance threshold δ and the weight coefficients of each loss λ_{coll}, $\lambda_{suppotr}$ and λ_{nav}, are determined through extensive grid search on the reserved validation set. Our optimisation objective is to maximise the rationality of the layout (high IoU) and physical consistency (low collision volume, high stability) in a coordinated manner. The search process reveals that the collision loss needs to be assigned a relatively high weight (λ_{coll} = 1.0) to effectively suppress the penetration phenomenon, while the support and passability loss, as supplementary constraints, have relatively low weights. The final selected parameter combination achieved the best balance of the above goals on the validation set. It is worth noting that when the parameters vary within a certain range, the model's performance shows robustness. However, extreme values can cause a certain constraint to be too strong or too weak, thereby compromising the generation quality.

3.5 Consistency simulation emulation verification

In order to quantitatively assess the physical consistency of the generated scenes, we have established an automated simulation verification process. The generated scene layout is converted to unified robot description format (URDF) or simulation description format (SDF) format and imported into the PyBullet physics simulation engine. In the simulation environment, we perform two key tests:

- Static stability test: a small random perturbation is applied to each object in the scene and the simulation is run for a number of steps. The stability score of an object, S_{stable}, is defined as the proportion of objects that do not experience significant displacement after the perturbation.
- Navigability test: randomly generate a start and end point in the scene and try to plan a path using a standard navigational intelligence (e.g., an A-algorithm or a reinforcement learning intelligence). The navigation success rate S_{nav} is defined as the proportion of start-end pairs that successfully plan a path.

These simulation metrics S_{stable} and S_{nav} are not used as losses during training, but rather as the final evaluation of the model's performance to demonstrate the superiority of this paper's approach in generating 'simulatable' environments.

4 Experimental validation

To systematically evaluate the effectiveness of the PCM-Transformer framework proposed in this paper, we conducted comprehensive experiments on the publicly available dataset Matterport3D. This chapter aims to answer three core questions through quantitative and qualitative analyses:

1 Does PCM-transformer outperform existing state-of-the-art methods in terms of visual and layout quality of scene generation?

2 Does the method offer significant advantages in ensuring the physical soundness and simulability of the generated scenes?

3 How do the key components of the model (3D prior, physical loss) each contribute to the final performance?

4.1 Experimental setup

- Dataset and pre-processing: this experiment use the Matterport3D dataset, which contains 90 high-quality 3D scans of indoor scenes. We randomly divide the training, validation, and test sets according to the ratio of 70%/10%/20% to ensure the diversity of scene types. For each scene, we extract the bounding boxes of all objects from it, including their category labels, sizes, positions and orientations. At the same time, we manually write a corresponding text description for each scene, e.g., 'a bedroom with a double bed, two nightstands and a closet'. In the pre-processing stage, we filtered out small (e.g., decorations) and infrequently occurring object categories, and finally retained 27 common furniture categories, and all scenes were normalised to a uniform coordinate system.

To achieve efficient training and focus on the main furniture layout, we have set clear filtering criteria:

1 Volume filtering: exclude objects with a bounding box volume less than (such as' cups', 'books',' table lamps'), as they have a minor impact on the macro layout and will increase the complexity of the scene.

2 Frequency filtering: exclude object categories that appear less than 20 times in the entire dataset (such as' bathtub ', 'piano', 'toilet') to ensure that the model has sufficient data samples to learn the reasonable placement of such objects.

After filtering, we have retained 27 of the most common furniture categories that play a decisive role in the structure and function of the space, such as' beds', 'sofas',' tables', 'chairs',' cabinets', etc. Our evaluation metrics are divided into two main categories to fully measure the quality of the generated scenes.

- Generation quality metrics
- Fréchet inception distance (FID).

We evaluate visual fidelity by rendering 2D images of the generated and real scenes from a specific viewpoint and computing the FID score between the distributions of features of these images. Lower FID values represent better visual quality.

- Category-averaged IoU: evaluates the similarity of layouts by calculating the IoU between the generated scene and the real scene on the object bounding box and averaging them by category.
- Physical reasonability indicator: collision volume (collision volume): calculate the total volume of all objects in the scene that penetrate each other (unit: cubic metre), the smaller the value, the better.
- Furniture stability score (FSS): calculate the proportion of furniture that has not moved (displacement > 0.1 metres) after applying a small perturbation to each piece of furniture in the simulation environment.
- Navigation success rate: 100 randomly generated origin-destination pairs in the scene are used for path planning using Algorithm A to calculate the percentage of successful arrival at the destination.

In terms of baseline models and implementation details, we choose three strong representative baseline models for comparison:

- ATISS: the autoregressive transformer model proposed by Paschalidou et al. (2021), which is the current state-of-the-art approach for object-based layout generation.
- SceneFormer: a transformer-based scene generation model proposed by Wang et al. (2021), which also uses autoregression but incorporates contextual information from the scene graph.
- Pure transformer: a baseline that uses only the text encoder and the transformer decoder, and does not include the 3D a priori coding and physical constraint loss we propose.

Our PCM-transformer model uses the Adam optimiser with an initial learning rate of 1e-4 and a batch size of 32. The weight hyperparameters in the loss function are set to $\lambda_{coll} = 1.0$, $\lambda_{support} = 0.5$, $\lambda_{nav} = 0.2$. All experiments were done on a workstation equipped with an NVIDIA RTX 3090 GPU.

4.2 *Quantitative results and analysis*

We performed a comprehensive evaluation of all methods on the test set using the above metrics, and the quantitative results are summarised in Table 1.

Table 1 Comparison of quantitative results on the Matterport3D test set

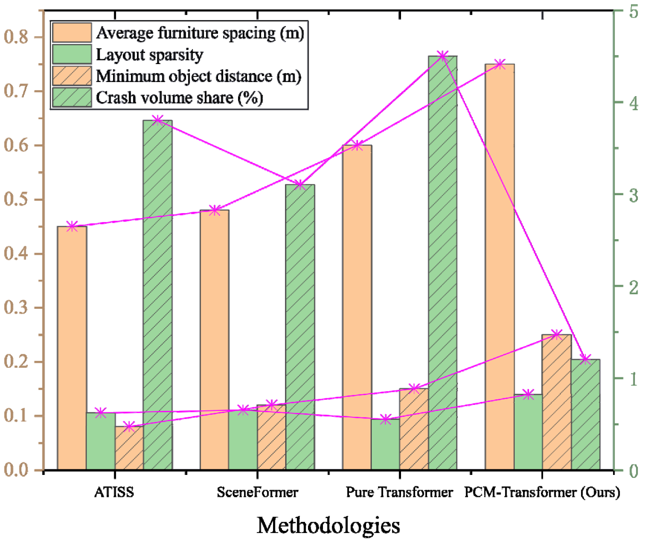
Methodologies	*FID*	*Category average IoU*	*Crash volume (m)*	*Stability score*	*Navigation success rate*
ATISS	45.2	0.412	0.38	0.81	0.72
SceneFormer	41.8	0.428	0.31	0.85	0.76
Pure transformer	50.1	0.395	0.45	0.75	0.68
PCM-transformer (Ours)	38.5	0.445	0.12	0.94	0.89

The following conclusions can be drawn from Table 1: First, in terms of generation quality, our PCM-transformer achieves the best performance in both FID and category IoU. This indicates that the introduction of a 3D scene prior and fusion through cross-modal attention effectively improves the visual fidelity and layout rationality of the generated scene. Secondly, and most critically, our method demonstrates an overwhelming advantage in the physical rationality metric. Compared to the optimal baseline, the collision volume drops by 61%, the FSS improves by about 10%, and the navigation success rate improves by 17%. This result strongly demonstrates the effectiveness of our proposed physical constraint loss function in guiding the model to generate structurally sound, simulation-ready environments. The pure transformer baseline is worse in all metrics, highlighting the limitations of relying only on text-layout mapping without explicit physical constraints and 3D a priori.

4.3 *Qualitative results and analysis*

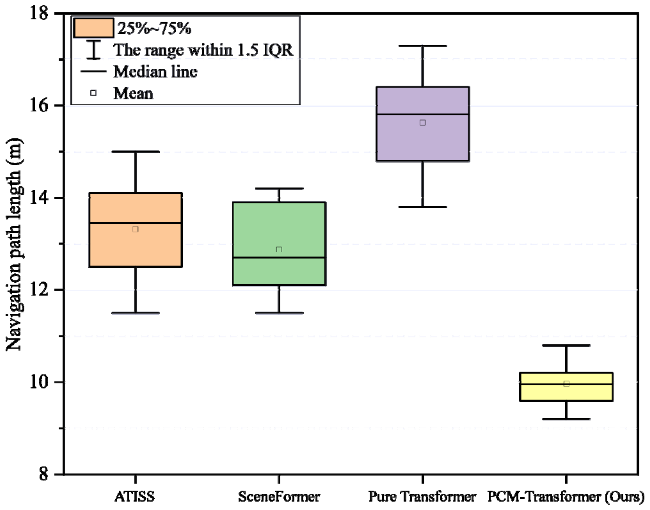
In order to objectively compare the generation quality of different methods, we quantitatively evaluate several key physical metrics of the generated scenes in Figure 2. By statistically analyzing a large number of scenes in the test set, it can be clearly observed that our PCM-transformer demonstrates significant advantages in all metrics. Specifically, in terms of average furniture spacing and layout sparsity, which measure layout rationality, our method achieves the highest score, indicating that the space utilisation of the generated scenes is more comfortable and reasonable. More importantly, our method far outperforms the baseline model in terms of minimum object distance and collision volume percentage, which directly reflect the physical consistency, proving the effectiveness of the physical constraint loss function, which is able to drastically reduce the penetration phenomenon between objects from the root.

Figure 2 Comparison of 2D layout and quantisation of spatial metrics of scenes generated by different methods (see online version for colours)



To further quantify the simulability of the generated scenarios, we count the performance of the virtual intelligences in 100 random navigation tasks in Figure 3. The box plots clearly show that in the scenarios generated by our method, the navigational path lengths of the intelligences are shorter and more centrally distributed, and the navigation success rate is much higher than that of the baseline method. This proves that the scenes generated by our method are structurally more conducive for the intelligent bodies to pass efficiently and reliably, whereas the scenes generated by the baseline method are not well laid out, resulting in the intelligent bodies needing to go around a longer distance or even failing the task.

Figure 3 Quantitative comparison of navigation performance of intelligentsia in scenes generated by different methods (see online version for colours)



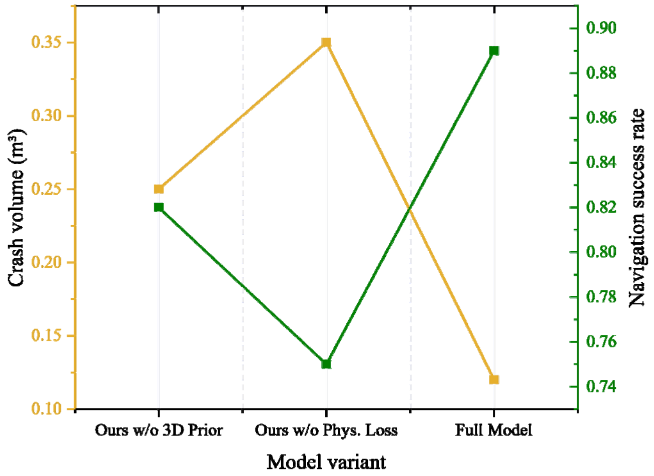
4.4 Ablation experiment

In order to dissect the contribution of each component in the model, we designed detailed ablation experiments, and the results are presented in the form of bar charts in Figure 4.

We constructed three variants of the model:

- Ours w/o 3D prior: removes the 3D scene prior encoder and cross-modal attention.
- Ours w/o Phys. loss: removes all physical constraint losses (L_{coll}, $L_{support}$, L_{nav}) and keeps only the reconstruction loss L_{recon}.
- Full model: the complete PCM-transformer model.

Figure 4 Analysis of ablation experiments: impact of key components on physical conformance metrics (see online version for colours)



We selected the two most representative physical metrics, collision volume and navigation success rate, to demonstrate. The results clearly show that the full model has the optimal performance. Removing the 3D prior leads to a rise in collision volume and a decrease in navigation success rate, demonstrating the importance of learning spatial common sense from real data. The removal of the physical constraint loss has the most significant impact, with a sharp increase in collision volume to a level comparable to the baseline model and a significant decrease in navigation success rate, which fully demonstrates that the physical loss function we designed is the key to achieving generative consistency.

4.5 Experimental results and analysis

In this study, we have made significant progress in the task of text-based 3D environment generation by introducing physical constraints and multimodal fusion mechanisms. Experimental results consistently show that our PCM-transformer model outperforms existing state-of-the-art methods in terms of visual fidelity, layout rationality, and especially physical consistency. Behind these results lie important theoretical insights and practical implications.

First, the core theoretical contribution of this work lies in the successful deep integration of symbolic physical knowledge with data-driven learning. Traditional generative models such as GANs or VAEs mainly rely on statistical laws in the data, and their 'consistency' is often limited to the visual or semantic level. Interaction networks such as those proposed by Battaglia et al. (2016), on the other hand, accurately simulate physical dynamics but are more for analysis than creation. This study bridges the gap between the two. By designing microscopic loss functions for physical constraints (e.g., collision loss, support loss), we translate human a priori common sense about the physical world (e.g., 'objects are impenetrable', 'furniture needs to be placed in a stable manner') into optimisation objectives that are understood by neural networks. This suggests that there is a ceiling for the purely data-driven paradigm in complex generative tasks, and that the introduction of symbolic, rule-based constraints is a key path to guide the model to break through this ceiling and generate content that not only looks good but also works well. This is in line with the research direction of combining rational reasoning and perceptual learning advocated by Janner et al. (2018).

Second, our cross-modal Transformer architecture demonstrates strong representation fusion capabilities. The model's ability to generate physically rational layouts is largely due to the cross-attention mechanism that allows textual semantic features to be retrieved and aligned from 3D a priori features enriched with real-world physical laws. This implies that the model may have spontaneously developed an implicit encoding of physical laws in its internal representations during training. For example, when the decoder generates a 'bookshelf', it may not only activate semantic associations with 'book' and 'wall', but also, through the fused features, associates It may not only activate the semantic association with 'book' and 'wall', but also associate the spatial and physical concepts such as 'load-bearing' and 'placed against the wall' through the fused features. This kind of multimodal alignment, which goes beyond pure semantics, is similar to the visual-verbal alignment implemented in CLIP by Radford et al. (2021), but extends it to the more challenging visual-verbal-physical three-dimensional space, providing a useful exploration for constructing AI systems with a more grounded world model.

On a practical level, the environments with a high degree of physical consistency generated in this study open up new possibilities for downstream applications. In the field of robot simulation and training, as emphasised by Savva et al. (2019), the realism of the simulation environment is crucial. Our approach can generate a large number of well-structured, navigable and diverse training scenarios on-demand, thus drastically reducing the reliance on expensive manual modelling or real-world data collection. In digital twins and building information modelling, designers can input natural language concepts to quickly obtain multiple preliminary layout scenarios that are physically compliant and fully functional, greatly improving design efficiency. In addition, in meta-universe and game development, the approach enables automated and intelligent generation of content while ensuring interactivity and immersion in the generated world.

Nevertheless, this study still has several limitations. First, the model currently focuses on static physical consistency and has not yet is dynamic physical interactions (e.g., opening doors, pushing and pulling drawers). Second, our approach relies on high-quality, densely labelled datasets such as Matterport3D, which are costly to collect and label. Finally, the complexity of currently generated scenes is still limited by the common object classes and relationships in the dataset, and the generative ability of the model may be limited for textual descriptions that contain anomalous or highly creative layouts.

Based on the research results and existing limitations in this paper, we plan to explore deeply in the following directions in our future work. First, we will break through the limitation of static consistency and work on modelling dynamic physical interactions. Second, to reduce the reliance on high-cost labelled data such as Matterport3D, we will explore weakly-supervised and self-supervised learning paradigms. Finally, we will focus on improving the model's combinatorial generalisation and creative reasoning capabilities. In addition, multi-level environment generation that extends the scope of generation from the room scale to the whole building scale will also be an important research direction.

5 Conclusions

In this paper, we propose a novel generative framework driven by a multimodal Transformer to address the long-standing physical inconsistency problem in automated environment generation. The core innovation of the framework lies in a cross-modal fusion architecture that deeply integrates the semantic information of textual descriptions with the geometric and physical a priori learned from real 3D scenes, and introduces a series of explicit and microscopic physical constraint loss functions during the training process, so as to guide the model to generate not only visually realistic, semantically relevant, but also physically reasonable 3D environments that can be directly used in simulation. 3D environments.

We perform full experimental validation on the authoritative Matterport3D dataset. Quantitative and qualitative results consistently show that the PCM-Transformer model proposed in this paper reaches the state-of-the-art in standard generation quality metrics and significantly outperforms existing baseline methods in key physically plausible metrics (e.g., collision volume, stability, navigability). Systematic ablation experiments further confirm the indispensable role of 3D scene a priori knowledge and physical constraint loss in achieving generation consistency.

In summary, the main contribution of this study is to successfully embed the rules of the physical world in a computable form into a data-driven generative model, which promotes a paradigm shift from 'visual synthesis' to 'functional creation' in environment generation technology. This work not only provides a feasible technical solution for building the next generation of reliable AIGC tools, but also provides new ideas and insights for AI research in understanding and creating complex physical worlds. In the future, we will continue our in-depth exploration along the directions of dynamic interaction modelling, weakly supervised learning and improving generalisation ability.

Acknowledgements

This work is supported by the 2023 annual project of the '14th Five-Year Plan' for Education and Science in Guangxi (No. 2023B379).

Declarations

All authors declare that they have no conflicts of interest.

References

Alatan, A.A., Yemez, Y., Gudukbay, U., Zabulis, X., Muller, K., Erdem, Ç.E., Weigel, C. and Smolic, A. (2007) 'Scene representation technologies for 3DTV – a survey', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 17, No. 11, pp.1587–1605.

Battaglia, P., Pascanu, R., Lai, M. and Jimenez Rezende, D. (2016) 'Interaction networks for learning about objects, relations and physics', *Advances in Neural Information Processing Systems*, Vol. 29, p.621.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y. (2017) 'Matterport3D: learning from RGB-D data in indoor environments', *IEEE Computer Society*, Vol. 9, pp.667–676.

Coumans, E. and Bai, Y. (2016) 'Pybullet, a python module for physics simulation for games, robotics and machine learning', *Pybullet*, Vol. 1, p.1.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T. and Nießner, M. (2017) 'Scannet: richly-annotated 3D reconstructions of indoor scenes', *Computer Vision and Pattern Recognition*, Vol. 6, pp.5828–5839.

de Avila Belbute-Peres, F., Smith, K., Allen, K., Tenenbaum, J. and Kolter, J.Z. (2018) 'End-to-end differentiable physics for learning and control', *Advances in Neural Information Processing Systems*, Vol. 31, p.265.

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.665.

Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P. and Poole, B. (2022) 'Zero-shot text-guided object generation with dream fields', *Computer Vision and Pattern Recognition*, Vol. 13, pp.867–876.

Janner, M., Levine, S., Freeman, W.T., Tenenbaum, J.B., Finn, C. and Wu, J. (2018) 'Reasoning about physical interactions with object-oriented prediction and planning', *Learning Representations*, Vol. 12, p.10902.

Kingma, D.P. and Welling, M. (2013) 'Auto-encoding variational bayes', *Learning Representations*, Vol. 13, p.6114.

Li, M., Patil, A.G., Xu, K., Chaudhuri, S., Khan, O., Shamir, A., Tu, C., Chen, B., Cohen-Or, D. and Zhang, H. (2019) 'Grains: Generative recursive autoencoders for indoor scenes', *ACM Transactions on Graphics* (*TOG*), Vol. 38, No. 2, pp.1–16.

Lopez, R., Regier, J., Jordan, M.I. and Yosef, N. (2018) 'Information constraints on auto-encoding variational Bayes', *Advances in Neural Information Processing Systems*, Vol. 31, p.161.

Paschalidou, D., Kar, A., Shugrina, M., Kreis, K., Geiger, A. and Fidler, S. (2021) 'Atiss: autoregressive transformers for indoor scene synthesis', *Advances in Neural Information Processing Systems*, Vol. 34, pp.12013–12026.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P. and Clark, J. (2021) 'Learning transferable visual models from natural language supervision', *Machine Learning*, Vol. 16, pp.8748–8763.

Ritchie, D., Wang, K. and Lin, Y.-a. (2019) 'Fast and flexible indoor scene synthesis via deep convolutional generative models', *Computer Vision and Pattern Recognition*, Vol. 21, pp.6182–6190.

Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.-Y., Fumero, M. and Malekshan, K.R. (2022) 'Clip-forge: Towards zero-shot text-to-shape generation', *Computer Vision and Pattern Recognition*, Vol. 52, pp.18603–18613.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V. and Malik, J. (2019) 'Habitat: a platform for embodied AI research', *Computer Vision*, Vol. 12, pp.9339–9347.

Song, S., Lichtenberg, S.P. and Xiao, J. (2015) 'Sun RGB-D: a RGB-D scene understanding benchmark suite', *Computer Vision and Pattern Recognition*, Vol. 7, pp.567–576.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30, p.7941.

Wang, X., Yeshwanth, C. and Nießner, M. (2021) 'Sceneformer: indoor scene generation with transformers', *IEEE 3D Vision*, Vol. 6, pp.106–115.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R. and Ren, D. (2020) 'Distance-IoU loss: faster and better learning for bounding box regression', *Artificial Intelligence*, Vol. 34, No. 7, pp.12993–13000.