INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of
Environment and
Sustainable Development

International Journal of Environment and Sustainable Development

ISSN online: 1478-7466 - ISSN print: 1474-6778
https://www.inderscience.com/ijesd

Application of deep learning algorithms in the design of urban subway public art space

Qian Wang

DOI: 10.1504/IJESD.2025.10074930

Article History:

Received:	23 May 2025
Last revised:	21 August 2025
Accepted:	20 October 2025
Published online:	23 February 2026

Application of deep learning algorithms in the design of urban subway public art space

Qian Wang

Faculty of Humanities and Arts,
Macau University of Science and Technology,
Macau, China
Email: hyhwqhw@163.com

Abstract: This paper aims to address the problem of insufficient integration of user visual attention behaviour modelling and spatial practicality in the design of subway public art spaces. This paper first constructs a cultural semantic labelling system and spatial attribute structure for subway stations based on deep neural networks. Second, it achieves multimodal deep alignment between semantic content and visual images through a contrastive language-image pre-training (CLIP) model. Then, it designs a multi-objective optimisation generation framework. Finally, it introduces a spatial structure adaptive analysis mechanism to achieve deep integration of generated content with the real subway environment. Experimental results show that each cultural category achieved high average scores of 0.9025 and 0.88 in terms of semantic consistency and accuracy of cultural background expression, respectively, indicating that the method performs well in terms of visual guidance effect and actual spatial foothold.

Keywords: visual attention modelling; public art space; deep learning; cultural semantic framework; spatial adaptability analysis.

Reference to this paper should be made as follows: Wang. Q. (2026) ‘Application of deep learning algorithms in the design of urban subway public art space’, *Int. J. Environment and Sustainable Development*, Vol. 25, No. 5, pp.44–72.

Biographical notes: Qian Wang is an Associate Professor currently pursuing a PhD. With a long-term commitment to environmental design and display design, she has extensive experience in teaching and research. Her academic achievements include publishing over ten professional papers and a 200,000-word academic monograph. She has led or participated in nine provincial and municipal-level research projects and received 20 high-level professional awards, including four national and 12 provincial recognitions

1 Introduction

As deep learning technology continues to break through the capabilities of image generation and semantic understanding, its potential for application in urban public art design is receiving widespread attention. As urban rail transit systems continue to expand, subway space not only carries transportation functions, but also gradually becomes an important venue for urban cultural communication and visual art presentation. As a frequently used public space, the environmental experience of subway stations has a significant impact on the public's aesthetic cognition (Xu and Zhou, 2022), cultural identity, and daily psychological state. The design of subway public art spaces (He and Gyergyak., 2021; Cheung et al., 2021) has shifted from single aesthetic decoration (Zhuoyu, 2023) to a multi-dimensional integration of cultural expression (Meng, 2023; Benthien, 2021), city image communication (Campos and Barbio, 2021; Matthews and Gadaloff, 2022), and citizen interactive experience (Milne and Pojani, 2023). Its design strategy (Shang and Halabi, 2024) needs to take into account functionality, safety, culture, and artistry.

However, current public art design is still mainly based on manual creation, with a long design cycle, limited content expression and a lack of systematic summary of cultural connotations, making it difficult to meet the rapidly expanding needs of subway construction. As deep learning algorithms (Zhang, 2022; Castellano and Vessio, 2021) have made significant progress in the fields of image generation and style transfer (Pu and Li, 2023), the intelligentisation of design processes has gradually become a research hotspot, but the application of existing methods in the field of public art still faces many challenges. The most critical problem is that existing deep generative models (Suzuki and Matsuo, 2022; Ruthotto and Haber, 2021) are mainly driven by visual image features (Wang and Kim, 2022), and their ability to understand cultural semantics (Kumar et al., 2022; Lamas et al., 2024) is weak. The generated results often show a variety of styles but lack of urban cultural consistency.

In addition, the algorithm generation process lacks automatic judgment of spatial structure adaptability, and the feasibility of the generated image in the physical space is difficult to guarantee, which in turn limits its application depth in actual subway design projects. Insufficient semantic understanding ability directly affects the cultural expression quality of the generated results, while insufficient spatial adaptability affects the actual deployment efficiency of the design results (Yang, 2023). Therefore, how to improve the semantic control ability of the deep learning model and enhance its coupling mechanism with the physical space structure has become a core issue that needs to be overcome when introducing deep learning technology into the design of urban subway public art spaces.

This study proposes a deep learning-based intelligent generation method for subway public art space, which systematically solves key problems such as cultural semantic expression, visual attention simulation, and spatial deployment adaptation. The method first constructs a structured semantic embedding space based on the knowledge graph and models the spatial features in combination with the physical attributes of the station. Then, the CLIP model is used to realise the multimodal mapping between semantic labels and visual images to obtain image features with cultural expression orientation. On this basis, the visual attention behaviour of passengers is simulated by integrating the attention prediction network of image content, spatial layout, and crowd density. A multi-objective optimisation framework is introduced in the image generation process to ensure semantic consistency, rationality of visual focus, and spatial matching, and a differentiable optimisation method is used to control the key structure of the generated image. To achieve the deployability of the generated results, a spatial topology adaptation module based on conditional GAN is designed to embed the image into the site structure model to evaluate the fusion effect and achieve a high degree of consistency between the content and the environment. The overall approach constructs a closed-loop path from semantic construction to image generation and deployment, provides a complete data-driven solution for the intelligent design of subway public art, and expands the research paradigm of cultural communication and visual guidance in high-traffic environments.

2 Related work

In the current research on the design of urban subway public art spaces, many scholars have explored its design concepts, strategies and practical applications from different perspectives. Lee (2022) analysed the Moscow Metro and pointed out that its spatial layout not only meets the basic transportation function, but also incorporates humanistic and social elements into the design, giving the public art space a deeper social meaning. This transition from function to culture provides important inspiration for subsequent research. Following this line of thought, Lian (2023) proposed a comprehensive design strategy from the macro, meso, and micro levels when designing a subway public art system based on the urban spirit, and supplemented it with a variety of artistic design element transformation methods, emphasising that subway space should be linked with urban culture. To enhance the user's interactive experience, Xinxin and Hashim (2024) analysed the integration path of public art and subway space from the perspective of interactive design and proposed a series of design strategies to optimise the interactive effect, so as to enhance the emotional connection between passengers and the environment. Although the above studies have put forward many useful ideas from the aspects of spatial function, humanistic connotation, and interactive experience, overall, there is still a lack of relevant research on how to guide the matching and expression of public art content and form by deeply mining the semantic information of cultural data (Sehar et al., 2021).

In recent years, deep learning technology has shown an increasingly broad application potential in solving practical problems in subway public art space design. Shi et al. (2025) combined deep learning sentiment computing with semantic segmentation methods and evaluated the correspondence between the physical characteristics of subway station space and user emotional perception under the theoretical framework of environmental psychology. They further revealed the key factors that affected passengers' psychological experience and provided data support for the emotion-oriented optimisation of space design. This demonstrated deep learning's potential for understanding and structuring human-computer relationships (Šumak et al., 2022) and also laid a technical foundation for emotional design (Chen, 2024). By constructing a dedicated dataset and improving the algorithm, I revealed the correlation between the visual content and emotional expression of the paintings, providing a new technical perspective for the study of traditional culture. These studies together show that deep learning can assist artistic expression and emotional communication, and also improve feedback efficiency (Kang et al., 2021; Menghani, 2023) and controllability in the design process. However, although relevant research has achieved certain results in perceptual data analysis and emotional response modelling, there are still bottlenecks in the lack of technical adaptation and algorithm perception (Lei et al, 2024) in effectively identifying and understanding the deep semantics implied in the cultural context. In the process of art generation, there is still a significant research gap in how to use deep learning to model user visual attention behaviour and then integrate it with the structural practical needs of subway space. On the one hand, existing methods mostly focus on image style generation or emotion recognition and lack a modelling mechanism for passenger visual focus patterns; on the other hand, the deployment of public art content often does not fully consider the coupling relationship between spatial structural constraints and visual guidance effects. This problem limits the ability of deep learning technology to generate art content with cultural semantics, visual appeal, and deployment feasibility in real environments. Therefore, it is urgent to build a deep generation framework that integrates semantic understanding, visual prediction and spatial adaptation to improve the intelligence level and practical implementation of subway public art design.

3 Design of multimodal semantic-driven spatial generation method

3.1 Cultural semantic embedding and spatial topology modelling of subway stations based on knowledge graphs

This section constructs a multi-dimensional cultural semantic label system, which aims to capture the cultural and historical background, urban context distribution, and artistic expression style characteristics of the area where the subway station is located. The system obtains cultural information from literature such as local chronicles, urban planning texts, and art review archives. By preprocessing these original texts, the Bidirectional Encoder Representations from Transformers (BERT) model is used to embed the texts, thereby obtaining the vectorised representation of each word. To improve the accuracy and representativeness of cultural semantic information, local chronicles, local cultural and historical materials, and urban planning texts are further introduced as the basis for text screening. Filtering criteria are formulated based on the authority and timeliness of published materials within the administrative jurisdiction to optimise the content basis of semantic annotation from the source.

To ensure that the semantic information in the text can accurately reflect the cultural background, the cosine similarity is used to calculate the similarity between the words in the text and construct a semantic map. The word vector representation of the text is obtained through the BERT model. Let v_i, $v_j \in R^d$ be the vector representation of two words in the BERT model, where d is the dimension of the word vector. The similarity between words is measured by calculating the cosine similarity, and the calculation formula is:

$$\text{sim}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \tag{1}$$

$v_i \cdot v_j$ is the inner product of word vectors v_i and v_j, and $\| v_i \|$ and $\| v_j \|$ are the norms of their respective vectors. This formula is used to measure the semantic relevance between two word vectors. The larger the value, the more similar the semantics between the words.

By calculating the similarity between all word pairs in the document, a high-dimensional similarity matrix $S \in \mathbb{R}^{n \times n}$ is constructed, where n is the size of the vocabulary. Based on this similarity matrix, the spectral clustering algorithm is used to cluster the words, and several word groups with strong semantic relevance are obtained, corresponding to different cultural feature labels in the cultural background of subway stations.

Although spectral clustering can effectively identify static semantic groups, it is difficult to capture the dynamic associations between labels. To this end, this study proposes to further introduce time series analysis methods based on the existing graph convolutional network (GCN) modelling to detect the trend of label semantics evolving over time. For example, a timestamp label map is established based on the age information of cultural texts, and the evolution path of semantic labels is identified using a sliding window mechanism, thereby revealing the changing trajectory of different cultural themes in urban development. This helps improve the responsiveness of the semantic labelling system to dynamic contexts. This study introduces a GCN to model the high-order semantic relationship graph between labels. Specifically, historical tags and regional tags are used as graph nodes, and the edge weights are determined by the semantic association strength calculated by cosine similarity. GCN dynamically updates node embedding representations through a multi-layer neighbourhood information aggregation mechanism to capture high-order semantic associations between tags.

Based on the dynamic modelling of semantic relationships by GCN, to further optimise the construction of semantic tags, the term frequency-inverse document frequency (TF-IDF) weighting mechanism is introduced to enhance the importance of words in cultural contexts. The calculation formula of TF-IDF is:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{2}$$

TF(t, d) is the frequency of term t in document d, calculated as:

$$\text{TF}(t, d) = \frac{\text{count}(t, d)}{\sum_{t'} \text{count}(t', d)} \tag{3}$$

count(t, d) represents the number of occurrences of term t in document d, and $\sum_{t'} \text{count}(t', d)$ is the total number of all terms in document d. Frequently occurring words can have a larger TF value, indicating that the word may have a higher semantic importance in the current document.

IDF(t) is the inverse document frequency of term t in the entire corpus, calculated as:

$$\text{IDF}(t) = \log \frac{N}{\text{df}(t)} \tag{4}$$

N is the total number of documents in the corpus, and df(t) is the number of documents containing the term t. The role of IDF is to reduce the weight of common words that appear frequently in all documents, thereby highlighting those words that appear less frequently in the corpus but have specific cultural backgrounds or important semantics. Through the weighted calculation of TF-IDF, words with higher cultural semantic value can obtain higher weights when generating tags.

Based on the construction of the semantic tag system, to further build an attribute system integrating geographic spatial information, this study introduces the spatial location of subway stations as a supplementary dimension. Specifically, each site has a corresponding geographic coordinate, and its geographic attribute vector can be formalised as $\mathbf{p}_i$. To unify the scale of semantic vectors and spatial vectors, a standardisation method is used to normalise all geographic coordinates and concatenate them with semantic vectors to form a unified representation:

$$\mathbf{z}_i = [\mathbf{v}_i; \lambda \cdot \mathbf{p}_i] \tag{5}$$

Among them, λ is the scaling factor for adjusting the weight of spatial attributes, controlling the proportion of spatial information in the final clustering and labelling system. The concatenated vector $\mathbf{z}_i$ represents the comprehensive characteristics of the site in both semantic and spatial dimensions, so that the model can take into account both geographical proximity and cultural semantic relevance when measuring similarity and clustering.

After completing the deep modelling of cultural semantics, it is necessary to further integrate the physical space characteristics of the subway station and model the three-dimensional spatial attributes of the station. Considering the unstructured point cloud characteristics of the subway station CAD (Computer Aided Design) model, the point cloud neural network PointNet++ is used to process the station CAD model. The geometric data of the station CAD model is converted into point cloud format and input into the hierarchical feature extraction module of PointNet++. Then, through multi-scale grouping and multi-resolution sampling strategies, the network identifies spatial topological constraints from unstructured point clouds and outputs topological feature vectors aligned with the semantic label dimension.

The weighted term set can be used as the core element of the multimodal semantic label system and further reduced in dimension through UMAP (Uniform Manifold Approximation and Projection) to verify the stability of semantic clustering. The label vectors after UMAP dimensionality reduction can be presented in a visual way, which can intuitively observe the distribution and relationship of different labels in the cultural semantic space. Finally, all labels can be encoded into vector form to promote the smooth progress of the subsequent generation process, providing accurate and culturally rich semantic input for the subsequent image generation process. To enhance the clarity of cultural information organisation, a hierarchical structure model is constructed within the cultural semantic label system. This model divides cultural labels into multi-level nodes according to semantic granularity: the top level represents the broad cultural domain; the intermediate level represents sub-domains with shared thematic attributes; the bottom level represents fine-grained cultural elements with specific symbolic meanings. The hierarchical relationships are established by calculating semantic correlation strengths between labels and applying a directed acyclic graph structure to represent parent-child dependencies. This structure allows the semantic label space to not only preserve vector-level similarity but also display explicit organisational relationships, improving interpretability and guiding downstream multimodal alignment.

3.2 Semantic-visual feature alignment of multimodal CLIP-adapter

After completing the construction of the semantic label and spatial attribute system of subway stations, this paper further focuses on how to achieve alignment and understanding between multimodal semantics, especially to establish effective associations between images and texts. To this end, this paper introduces the CLIP model, which drives the unified modelling of semantic relationships through its powerful cross-modal alignment capabilities, and realises the semantic expression and alignment mechanism under visual guidance. To enhance the adaptability of the CLIP model to subway cultural semantics, an efficient parameter fine-tuning strategy based on Adapter is proposed. Specifically, a cultural semantic adaptation layer module is inserted between the Transformer layer of the visual encoder and text encoder of CLIP. The module consists of two layers of fully connected networks. The input dimension is consistent with the embedding dimension of CLIP, and the output is fused to the original CLIP features through residual connection. In addition, to further improve the accuracy and efficiency of semantic alignment, more flexible fusion strategies can be explored, such as introducing a gating mechanism to dynamically regulate the information flow between the adapter and the backbone network, or trying a cross-layer residual connection structure to strengthen the collaborative expression of high-level semantics and underlying features. At the same time, fusing graph neural networks or lightweight Transformer modules as an adaptation intermediate layer can also further optimise the depth and breadth of cross-modal alignment, providing a richer structural foundation for cultural semantic feature modelling.

Based on the cross-modal alignment mechanism of the contrastive language-image pre-trained CLIP model, a high-level semantic mapping between the cultural semantic labels of subway stations and the generated images is constructed by jointly optimising the embedding space of the text encoder and the image encoder. The structure design based on the CLIP model is shown in Figure 1.

Figure 1 Multimodal image-text fusion attribute recognition structure diagram based on CLIP (see online version for colours)

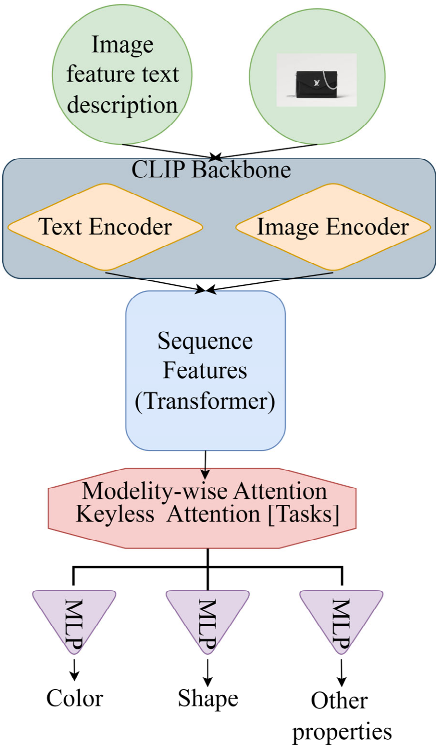


Figure 1 shows that the text encoder and image encoder in the CLIP model extract features of descriptive text and images, respectively, ensuring semantic consistency and enhancing the collaborative expression ability between modalities. Subsequently, the extracted sequence features are input into the Transformer module to strengthen the interaction of cross-modal information. This step improves the context relevance through the self-attention mechanism. After that, the keyless attention mechanism can be introduced to enable the model to dynamically focus on key modal features according to task categories without explicit guidance, effectively alleviating the problem of modal redundancy. Finally, different attribute tasks are output through independent multilayer perceptrons (MLP), ensuring the independence and non-interference of recognition tasks. The overall structural design is reasonable, ensuring the depth of multimodal feature fusion and the accuracy of attribute recognition.

To enhance the adaptability of the CLIP model in a multicultural context, the parameter fine-tuning strategy can be expanded through a more sophisticated adaptation mechanism while keeping the backbone structure frozen. On the one hand, gridding the hyperparameters of the adaptation layer (such as the intermediate dimension, the type of activation function, and the regularisation term) can effectively improve the stability of the model during the semantic transfer process; on the other hand, the introduction of a regional culture-driven initialisation method enables the model to have a perceptual bias towards specific cultural semantics at the beginning of training, thereby optimising the adaptation efficiency. In addition, for the fusion method between the visual encoder and the text encoder, a multi-stage or cross-attention-guided hierarchical adjustment strategy is adopted to help improve the alignment accuracy and structural expression ability of cultural entities.

For the specific aspects of the text encoder, it adopts the Transformer architecture, converts the input semantic label sequence $S = \{s1, s2, \ldots, sn\}$ into a dense vector through the word embedding function $Embed(\cdot)$, and passes it layer by layer to the l layer of the Transformer to generate the text feature representation $h_l^t = \text{Transformer}(Embed(S))$. The image encoder uses ResNet-50 (Residual Network 50 layers) as the backbone network, extracts the global feature vector h^i = ResNet-50(I) of the generated image, and maps it to the same dimension space as the text embedding through the linear projection layer $z^i = W_i h^i + b_{\mathrm{i}}$, where $W_i \in \mathbb{R}^{d \times d}$ is the learnable weight matrix, and d is the embedding dimension. The contrast loss function $\mathcal{L}_{\text{CLIP}}$ optimises the cosine similarity distribution of text-image embedding, and its mathematical form is:

$$\begin{aligned}\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum\nolimits_{i=1}^{N}\Bigg[&\log\frac{\exp\left(\text{sim}\left(z_i^t, z_i^i\right)/\tau\right)}{\sum\nolimits_{j=1}^{N}\exp\left(\text{sim}\left(z_i^t, z_j^i\right)/\tau\right)}\\ &+\log\frac{\exp\left(\text{sim}\left(z_i^i, z_i^t\right)/\tau\right)}{\sum\nolimits_{j=1}^{N}\exp\left(\text{sim}\left(z_i^i, z_j^t\right)/\tau\right)}\Bigg]\end{aligned} \tag{6}$$

z_i^t is the text embedding of the ith sample (corresponding to the ith cultural theme label sequence); z_i^i is the image embedding of the ith sample (corresponding to the ith generated image); z_j^i is the image embedding ($j \neq i$) of the jth sample, constituting the negative sample set of text z_i^t; z_j^t is the text embedding ($j \neq i$) of the jth sample, constituting the negative sample set of image z_i^i; τ is a temperature parameter that controls the sharpness of the similarity distribution and ensures the robustness of semantic alignment; N is the number of batch samples, which is set based on the trade-off between video memory capacity and training stability. This loss function forces the model to learn discriminative cross-modal representations by maximising the similarity of positive sample pairs and minimising the similarity of negative sample pairs. To further improve convergence speed and performance in specific cultural contexts, an initialisation strategy is introduced that aligns the Adapter parameters with cultural domain priors derived from pre-trained cultural semantic embeddings. The visual encoder and text encoder are pre-trained on a curated cultural dataset before integrating with the Adapter, ensuring that both modalities share a consistent cultural feature basis. During fine-tuning, the initialised parameters provide a stable starting point in the optimisation landscape, reducing the number of iterations required for convergence while preserving alignment accuracy.

The aligned text embedding z^t is dynamically associated with the denoising process of the diffusion model through the cross-attention mechanism. Specifically, in the t step denoising stage of the diffusion model, the text z^t embedding is linearly transformed to generate a query vector $Q_t = W_q z^t (W_q \in \mathbb{R}^{512\times512})$, and the attention weight is calculated with the key vector $K_t \in \mathbb{R}^{H\times W\times512}$ of the image feature map (H, W are the height and width of the feature map, extracted by the convolutional neural network):

$$\alpha_t = \text{softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) \tag{7}$$

Among them, $d_k = 64$ is the key vector dimension (divided by the number of feature map channels), and the scaling factor $\sqrt{d_k}$ is used to prevent the gradient from being unstable due to the dot product result being too large (for example, the magnitude of $Q_t K_t^T$ increases linearly with the increase of d_k). The attention weight α_t is multiplied by the value vector V_t to generate the semantically enhanced feature representation $F_t = \alpha_t V_t$, thereby dynamically integrating cultural semantic information in each diffusion process.

To quantify the semantic alignment effect, the cosine similarity mean μ_{sim} and standard deviation σ_{sim} are used as evaluation indicators:

$$\mu_{\text{sim}} = \frac{1}{M}\sum_{m=1}^{M} \text{sim}\left(z_{t_m}^t, z_{i_m}^i\right), \sigma_{\text{sim}} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\text{sim}\left(z_{t_m}^t, z_{i_m}^i\right) - \mu_{\text{sim}}\right)^2} \tag{8}$$

M is the number of test samples, and $z_{t_m}^t$, $z_{i_m}^i$ are the text embedding and image embedding of the mth sample. The specific evaluation results show that after the introduction of the Adapter fine-tuning module, the average cosine similarity between image and text embeddings increases from 0.842 to 0.902, and the standard deviation decreases from 0.061 to 0.045, indicating that the coupling between cross-modal features is significantly enhanced and more consistent. Furthermore, the top-1 image-text matching accuracy is used as a verification indicator, and the matching accuracy is increased from 78.6% to 88.2%, verifying the actual effect of the cultural adaptation layer in improving the accuracy of semantic alignment.

3.3 Visual attention prediction based on spatiotemporal transformer

After achieving preliminary semantic alignment with the CLIP model, the paper further focuses on the interests, preferences, and concerns shown by users in real interactions. To improve the model's ability to understand user behaviour patterns, the next step can be to combine user behaviour data and introduce an attention prediction network to explore the user's potential attention mechanism for image and text content.

In terms of network structure, a hybrid model architecture that integrates ViT (Vision Transformer) and TimeSformer is used to process subway surveillance video streams to predict passenger gaze heat maps. The ViT module is used to capture spatial visual features, and TimeSformer is responsible for modelling the temporal attention dependency between frames. This structure has a strong time-space feature modelling capability and can efficiently capture passengers' gaze trends and dynamic attention areas. In the Transformer module, a split attention mechanism is introduced to extract multi-scale dynamic features in parallel, and residual connections and layer normalisation are combined to improve training stability. To further improve the fine-grained expression capability of visual behaviour simulation, the model can try to introduce more advanced attention mechanisms in the future, such as self-attention to strengthen the modelling of internal correlations of features, or cross-attention to enhance the interactive coupling between multimodal inputs. These mechanisms can more accurately capture the relationship between local saliency and overall context in the process of predicting user visual behaviour, and improve the prediction quality and spatial perception ability of the attention map. The final output is a two-dimensional heat map, indicating the visual attention probability of different spatial positions.

In the modelling process of the user attention prediction network, a convolutional neural network combined with the attention mechanism is used to construct a simulation system for passenger visual focus behaviour. The network structure adopts a dual-path visual encoding-decoding framework with U-Net as the backbone. In the encoding path, the image input is processed by multi-layer convolution, batch normalisation, and ReLU activation function to extract multi-scale spatial hierarchical features; in the decoding path, the spatial resolution is restored layer by layer through deconvolution, and high-level semantics and low-level detail features are integrated. To enhance the model's response sensitivity to visual focus, a Squeeze-and-Excitation (SE) channel attention module is introduced in each jump connection to reweight the inter-channel features to strengthen the feature channel response that is highly correlated with the hotspot of human eye gaze.

In terms of input data, the model uses the video stream from the surveillance camera in the station area as the main visual input. The video format is RGB; the resolution is unified to 1920 × 1080 pixels; the frame rate is 25fps. At the same time, the subway gate pass data is introduced as a time synchronisation reference. The gate record includes the pass timestamp, pass direction, and anonymous identification of the passenger identity, which is used to assist in locating the behaviour nodes of users entering and leaving the area. By temporally aligning the video frames with the gate event sequence, the joint modelling of behavioural events and gaze data is achieved, providing a high-temporal and spatial resolution data basis for heat map prediction.

To simulate the real visual behaviour of users in complex subway space scenes, the real gaze point dataset is introduced as a supervisory signal in the network training stage. Taking the gaze heat map of each input image as the target, the KL divergence loss function is used to measure the deviation between the predicted probability distribution and the true gaze distribution. Assuming the image size is $H \times W$; (x, y) represents the pixel position of the x row and the y column in the image; $P(x, y)$ represents the probability value of the true gaze map at this position; $\hat{P}(x, y)$ represents the predicted probability value output by the model, then the KL divergence loss function is defined as:

$$\mathcal{L}_{KL} = \sum_{x=1}^{H} \sum_{y=1}^{W} P(x, y) \log\left(\frac{P(x, y)}{\overline{\hat{P}}(x, y)} \right) \tag{9}$$

To further enhance the overall trend consistency between the predicted image and the true attentive image, the Pearson correlation coefficient is introduced as an auxiliary loss term. Among them, P and $\overline{\hat{P}}$ represent the average values of the true attentive image $P(x, y)$ and the predicted image $\hat{P}(x, y)$, respectively. The Pearson correlation coefficient loss function is defined as:

$$\mathcal{L}_{CC} = -\frac{\sum_{x=1}^{H} \sum_{y=1}^{W} \left(\hat{P}(x, y) - \overline{\hat{P}} \right) \left(P(x, y) - \overline{P} \right)}{\sqrt{\sum_{x=1}^{H} \sum_{y=1}^{W} \left(\hat{P}(x, y) - \overline{\hat{P}} \right)^2} \cdot \sqrt{\sum_{x=1}^{H} \sum_{y=1}^{W} \left(P(x, y) - \overline{P} \right)^2}} \tag{10}$$

The final comprehensive loss function is constructed in the form of a weighted combination to optimise both the local pixel-level distribution differences and the overall statistical correlation:

$$\mathcal{L} = \alpha \mathcal{L}_{KL} + \beta \mathcal{L}_{CC} \tag{11}$$

Among them, $\alpha = 0.7$, $\beta = 0.3$. Experiments have verified that this ratio has a better balance between improving model stability and prediction accuracy.

To enhance the model's ability to simulate actual user behaviour, on the basis of the original visual gaze data, behavioural dimensions such as dwell time distribution, movement path trajectory, and transfer mode between sites are introduced as auxiliary variables to participate in the predictive modelling of attention maps. This type of data has stability and continuity in the temporal structure, and can provide prior behavioural cues in areas where visual cues are missing, especially in cultural scenes, where has a significant guiding effect. The information at the behavioural level is synchronously encoded with the visual trunk channel during the model training stage, so that the attention map not only reflects the user's visual interest points, but also reflects the cumulative impact of their movement behaviour.

In the pre-processing stage, the input image is unified to 256 × 256 pixel resolution, and Z-score normalisation is performed to standardise the pixel value distribution. To improve the generalisation performance of the model, data enhancement strategies are introduced, including random rotation, brightness perturbation, and random cropping, to construct a diverse training sample set.

The network output is a two-dimensional floating-point probability map with a value range normalised to [0, 1], which is used to represent the probability distribution of the user's visual attention at each pixel position in the image. This prediction map can serve as the spatial attention guidance input of the subsequent image generation model, dynamically regulating the spatial configuration density and detail intensity of visual resources during the synthesis of artistic images, and achieving explicit coupling and effective response between the generated content and the visual focusing behaviour of the human eye.

To reveal time-varying characteristics of user visual attention, the attention prediction network incorporates behavioural data segmented by different temporal periods such as weekdays and weekends. The video stream and gate passage records are labelled with temporal attributes, and separate attention maps are generated for each temporal category. Comparative analysis between these categories highlights variations in hotspot distribution, fixation duration, and gaze shift frequency across different operation cycles, enriching the behavioural modelling dimension and supporting context-aware content generation.

3.4 Pareto optimal diffusion model generation framework

Based on modelling user attention, the paper realises that the optimisation method of a single objective can no longer meet the multi-dimensional requirements of expression quality, semantic consistency, and user preferences in multimodal systems. Therefore, the paper designs a generation framework that integrates multiple optimisation objectives to achieve a more comprehensive model performance improvement. The image generation process uses the latent diffusion model (LDM) architecture to embed the image generation task into the latent space for efficient optimisation. The conditional input of the model includes semantic embedding, predicted attention heatmap, and spatial topological feature vector extracted by PointNet++, realising the deep fusion of multi-source heterogeneous conditions.

In the multi-objective optimisation generation framework, the image generation task is defined as an optimisation problem under the three constraints of cultural expression, spatial functional adaptability, and user visual attention. The process is based on the generative adversarial network (GAN). The generator accepts the fused cultural semantic vector, spatial attribute vector, and visual attention probability map as input. The discriminator combines the image and semantic label to jointly evaluate the authenticity and semantic consistency of the generated image. The optimisation objective function is defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{sem} + \lambda_3 \mathcal{L}_{spa} + \lambda_4 \mathcal{L}_{att} \tag{12}$$

$\mathcal{L}_{adv}$ is the adversarial loss term, which is modelled using the minimum-maximum game objective in the standard GAN structure:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{I}_{real}}\left[\log D\left(\mathbf{I}_{real}, \mathbf{s}_c\right)\right] + E_{\mathbf{I}_g}\left[\log\left(1 - D\left(\mathbf{I}_g, \mathbf{s}_c\right)\right)\right] \tag{13}$$

$\mathcal{L}_{sem}$ is the semantic consistency loss, which is constructed through the image-text similarity measure of the CLIP model. Its goal is to maximise the multimodal similarity between the generated image and the input semantics:

$$\mathcal{L}_{sem} = 1 - \cos\left(\phi_{img}\left(\mathbf{I}_g\right), \phi_{txt}\left(\mathbf{s}_c\right)\right) \tag{14}$$

ϕ_{img} and ϕ_{txt} are the image and text encoders in the CLIP model, respectively, and $\cos(\cdot,\cdot)$ is the cosine similarity function.

$\mathcal{L}_{spa}$ is the spatial structure adaptation loss, which is calculated based on the matching degree between the deployable area mask $\mathbf{M}_{\underline{\text{deploy}}}$ in the three-dimensional space model and the image projection area $\mathbf{P}_{proj}$:

$$\mathcal{L}_{spa} = 1 - \frac{\left|\mathbf{M}_{deploy} \cap \mathbf{P}_{proj}\right|}{\left|\mathbf{P}_{proj}\right|} \tag{15}$$

This item is used to penalise the area ratio of the non-deployable part in the generated image to ensure that the image content is consistent with the spatial visible area.

$\mathcal{L}_{att}$ is the visual attention guided loss. The Kullback-Leibler (KL) divergence between the predicted visual attention probability map $\mathbf{A}_v$ and the image saliency map $\mathbf{S}_g$ is introduced for modelling to promote the generated content to be concentrated in the high attention area:

$$\mathcal{L}_{att} = \sum\nolimits_{i,j} \mathbf{A}_v(i, j) \log\left(\frac{\mathbf{A}_v(i, j)}{\mathbf{S}_g(i, j)}\right) \tag{16}$$

$\mathbf{S}_g$ is extracted from the generated image through the Saliency network. During the optimisation process, the generator parameters and the discriminator parameters are updated alternately, and the Adam optimiser is used to iteratively minimise $\mathcal{L}_{total}$ and improve the multi-objective adaptation capability. The loss weights λ_1, λ_2, λ_3, λ_4 are determined through a preliminary grid search to balance the contribution of each optimisation target, so that the final generated image can achieve the overall optimal performance solution in terms of semantic accuracy, spatial deployment feasibility, and visual appeal. To prevent overfitting to the training data, an adversarial regularisation mechanism is integrated into the generation framework. The generator is periodically updated through an auxiliary discriminator trained to distinguish between generated content adapted to seen cultural-spatial contexts and unseen ones, encouraging generalisable feature learning. Additionally, spectral normalisation is applied to convolutional layers in the generator and discriminator to stabilise training and constrain parameter growth, ensuring that the optimisation process maintains robustness across diverse cultural and spatial inputs. To further improve the content diversity and social value expression ability of the generated results, the model can expand the types of objective functions with more cultural significance in the future, such as semantic discreteness indicators oriented towards cultural innovation, green pattern generation indicators oriented towards environmental sustainability, etc., and incorporate sociality and aesthetics into the optimisation framework side by side, so as to promote cultural expression from formal coordination to composite value optimisation of semantic and ecological integration. The entire training process is conducted jointly on the training set and the validation set to ensure that the generator has stable generalisation capabilities under different combinations of cultural semantics and spatial structures.

3.5 Spatial topology adaptation module based on conditional gan

The multi-objective optimisation framework can achieve collaborative learning at multiple semantic levels, but in the context of the complex spatial structure of subway stations, the model generation results must also be highly adapted to the real spatial layout. Therefore, the spatial structure adaptability analysis mechanism can be introduced to evaluate and optimise the application effect of the model in the urban space context. In the spatial structure adaptability analysis mechanism, a process integrating spatial geometry modelling, scale consistency calculation, traffic line conflict detection, and visible area matching evaluation is adopted to achieve accurate alignment between the generated image and the actual structure of the subway station and verify the feasibility of deployment. Firstly, a 3D subway station structure model is constructed. The parametric component data of the station hall and platform area is exported using the building information model. A 3D spatial constraint vector set is generated through point cloud reconstruction and parameter translation. Each component vector contains size, position, and visual attribute labels. To evaluate the actual deployment occlusion of the generated art images in the three-dimensional spatial structure of the site, this paper further introduces the cycle-consistent generative adversarial network (CycleGAN) network to realise the image domain mapping between the generated images and the UV mapping space of the site CAD model. CycleGAN can learn the bidirectional conversion function between natural image style and CAD texture style without relying on paired data, so as to seamlessly project the image content onto the surface of site components. By overlapping the mapping results with the UV texture mask of the BIM (Building Information modelling) component surface, the structural occlusion rate index can be calculated to measure the proportion of the deployed image that is occluded, providing a graphic-level geometric constraint metric for spatial adaptability. The deployment area of the generated image is mapped between the plane image coordinates and the three-dimensional space projection points by the position mapping function, and then the image coverage area is calculated.

The scale matching analysis is based on the joint implementation of the Hausdorff distance d_H and the intersection over union (IoU) similarity measurement. The boundary point set of the space occupied by the generated image is defined as B_g, and the boundary of the corresponding target deployment area component is B_s. Then, the scale matching measurement function is:

$$M_{\text{scale}} = \exp\left(-\alpha d_H\left(B_g, B_s\right)\right) \cdot \text{IoU}\left(B_g, B_s\right) \tag{17}$$

Among them, α is the scale adjustment factor; Hausdorff distance is used to characterise the maximum deviation between boundary points; IoU reflects the overlap ratio, and the product of the two reflects the comprehensive score of scale matching.

Traffic line conflict detection depends on the path connectivity graph $G = (V, E)$, where node V represents the key line anchor point, and edge E represents the passable path. In the generated image deployment area, a blocking area is constructed for all affected node sets, and the connected subgraph is recalculated. If $\forall u, v \in E$, $\text{path}_{G'}(u, v) = \varnothing$ exist, the path is judged to be broken, and the dynamic line coordination function is defined:

$$M_{\text{flow}} = 1 - \frac{|Z|}{\sum_{(u,v)\in E} 1_{path_{G'}(u,v)=\varnothing}} \tag{18}$$

$|Z|$ is the volume of the blocked area, and the denominator is the total number of broken path logarithms, reflecting the tolerable degree of the impact of image deployment on traffic lines.

The visual area matching evaluation uses the heat map generated from the user attention prediction network, which is mapped to the three-dimensional space through the projection function to form the visual focus area volume V_h. The target area of the generated image is A_g, and the visibility matching degree is defined as:

$$M_{\text{vis}} = \frac{\text{Vol}\left(A_g \cap V_h\right)}{\text{Vol}\left(A_g\right)} \tag{19}$$

Here, *Vol* represents the volume operator, which reflects the proportion of the generated image in the user's visible area. The higher the index, the stronger the visual guidance of the design deployment.

The final adaptability score is achieved through the weighted fusion of the three indicators. The adaptability function is defined as:

$$M_{\text{adapt}} = \lambda_1 M_{\text{scale}} + \lambda_2 M_{\text{flow}} + \lambda_3 M_{\text{vis}} \tag{20}$$

λ_1, λ_2, $\lambda_3 \in [0, 1]$, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$ are met. The weights are dynamically adjusted according to the design priority to achieve the comprehensive optimisation of the constraints of scale, traffic, and visibility, ensuring the structural fusion and deployment feasibility of the generated image in the real subway environment. The time variation characteristics of site passenger flow are further introduced into the deployment strategy evaluation, and the traffic distribution during peak hours and off-peak hours is used as a dynamic weight factor for topology deployability to reflect the changes in the acceptance rate and achievable efficiency of content deployment in different time windows. This mechanism can improve the responsiveness of the deployment strategy to the actual operation rhythm and enhance the applicability of the model in time-varying scenarios. Based on the above adaptability evaluation results, the system can eventually generate image content embedded in the three-dimensional BIM model structure and output an IFC format model file with deployment information. A facility conflict detection process is integrated into the spatial adaptability analysis to ensure compatibility with existing subway infrastructure such as billboards, signage, and lighting installations. The 3D station model incorporates precise location and size data of these facilities, and an overlap detection algorithm computes the intersection ratio between projected image deployment areas and facility boundaries. Content deployment plans are adjusted to maintain zero intersection with critical information carriers, ensuring that generated designs do not obstruct operational signage or safety-relevant displays. The BIM file includes metadata such as mapping coordinates, occlusion area annotations, and visual focus area projections to ensure that the image content can be directly integrated into the construction design or simulation platform to support subsequent engineering applications, construction rehearsals, and digital twin deployment.

4 Experimental design and results analysis

4.1 Dataset construction and preprocessing process

In the study of subway public art space design, the construction and preprocessing of the dataset is a key link, which determines the quality and effect of subsequent model training and evaluation. The construction of the dataset is based on the collection of cultural information related to subway stations, covering text resources such as local chronicles, urban planning materials, and art reviews. To verify the adaptability and robustness of the model in different cultural regional contexts, the dataset is expanded to multiple subway line areas with cultural diversity, covering typical stations in the southeast coast, inland historical and cultural areas, and ethnic minority areas. New samples are included in the graphic annotation system through unified collection standards for cross-regional migration of the model and label robustness experiments. By processing these text resources and using natural language processing technology, the paper extracts key words with cultural significance and measures the semantic relationship between words through cosine similarity. This process helps build a multi-dimensional cultural semantic labelling system and also provides high-quality semantic input for subsequent image generation tasks.

The data preprocessing stage standardises and cleans the collected image data to ensure the consistency between the image and the cultural label. The selection of image data follows specific cultural theme requirements, screening out artworks that meet design requirements while ensuring the standardisation of image resolution, size, and colour. In addition, data preprocessing also includes labelling input images so that subsequent models can accurately understand the relationship between image content and the cultural connotations behind it. The core of this process is to ensure that the input data is well representative and usable through label standardisation and image processing, providing a solid foundation for model training and reasoning.

The constructed and preprocessed dataset is the basis for generating art design images and also provides a guarantee for the efficient operation of the model during the training phase. Through these carefully designed processes, the model can better understand and process the cultural background of subway stations, improve the performance of generated images in semantic consistency, artistic style, and spatial adaptability, and ensure the feasibility of the design scheme in the actual subway public art space. Figure 2 shows the layout of the signage of different subway stations involved in the experiment.

Figure 2 Layout of signage for different subway stations (see online version for colours)

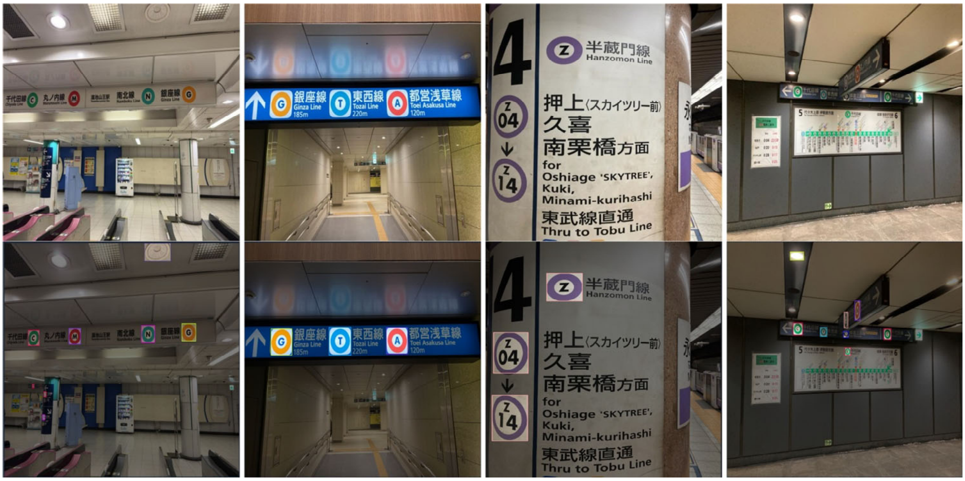


Figure 2 shows the sign system inside a subway station, showing different forms of signs inside the station, including directional signs, floor information, and transfer channel instructions. Comparing the sign settings in different areas, it can be seen that the clarity, colour contrast, and location layout of various information have a direct impact on passenger navigation. The design of the station signs uses different colours and graphic symbols to distinguish different functional areas, and at the same time magnifies the key information so that passengers can quickly identify it. However, the poor lighting conditions in some areas make the visual effect unclear, thus affecting the overall guidance effect. Through these analyses, it is possible to clearly point out the close relationship between the visual communication effect of the signs and the physical environment of the space, and how to achieve a more efficient guidance effect through design optimisation.

4.2 Model training and configuration details

The model training environment can be equipped with high-performance hardware and GPUs with powerful computing capabilities to meet the needs of deep learning models. The framework used is PyTorch 2.0, which is optimised for large-scale deep model training. To improve the reproducibility and engineering transparency of the model training process, this paper further clarifies the hyperparameter configuration of each core module and the hardware environment it relies on. The CLIP-Adapter module adopts a parameter-efficient fine-tuning strategy, with 20 training rounds, a learning rate of 5e-5, an AdamW optimiser, a weight decay parameter of 0.01, and a batch size of 64. In terms of the parameter freezing strategy, only the Adapter module and the final projection layer are fine-tuned, and the remaining Transformer layers remain frozen to ensure the stability of the semantic embedding structure.

In terms of diffusion model, the LDM architecture is used for latent space image generation; the diffusion step number is set to 1000; the linear beta noise scheduler is used (from 1e-4 to 0.02); the DDIM (Denoising Diffusion Implicit Models) sampler is used in the sampling stage; the sampling step number is set to 50; the latent space resolution is 4 × 64 × 64.

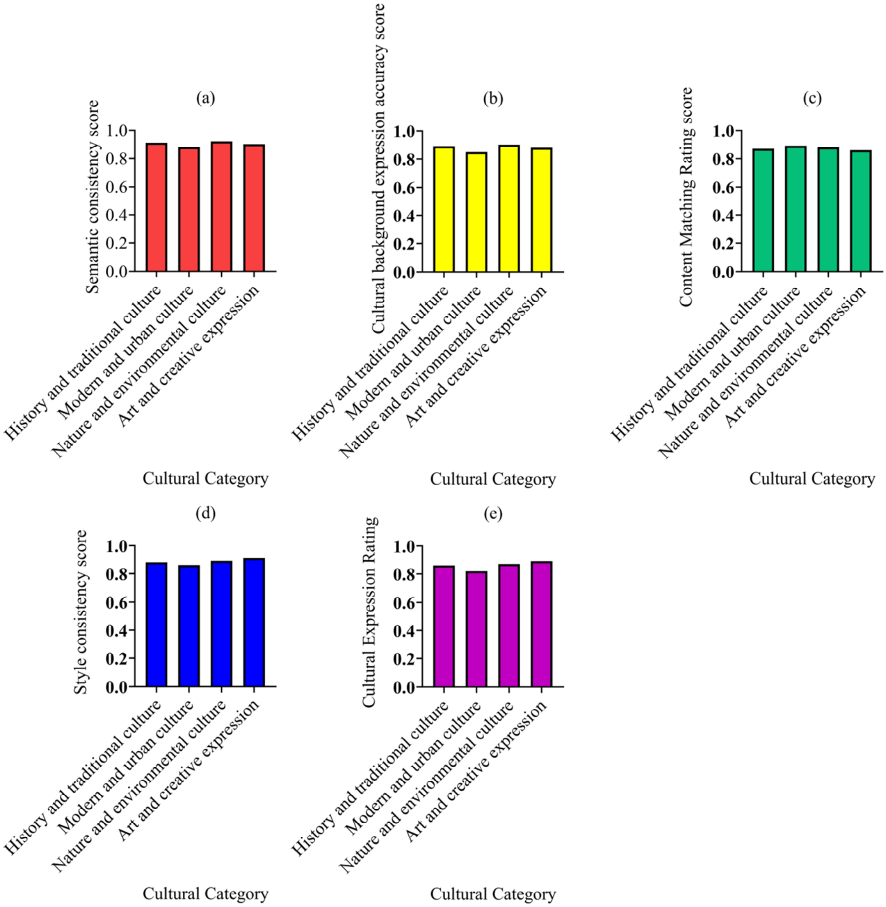
The training task is run on a high-performance computing platform with a NVIDIA A100 40GB GPU, an Intel Xeon Platinum 8352Y dual-core processor, and 512GB DDR4 ECC memory. The operating system is Ubuntu 22.04 LTS; the CUDA version is 11.7; the cuDNN version is 8.4; the deep learning framework is based on PyTorch 2.0.1 and HuggingFace Transformers 4.30. To evaluate the impact of hardware configuration differences on training efficiency and generation quality, this study completes model training under two sets of experimental conditions: Configuration A uses an NVIDIA RTX 3090 graphics card with 24GB of video memory; Configuration B uses an NVIDIA A100 graphics card with 40GB of video memory. The experiment shows that under conditions of higher video memory and bandwidth, the training cycle is shortened by an average of 21%, while the visual consistency index of the generated images is improved by 5.6%, showing better style integrity and spatial structure retention capabilities. This result provides a quantitative reference for the subsequent model deployment and performance requirement matching.

In terms of loss function design, cross entropy loss and regularisation are combined to minimise the error between the generated image and the target cultural semantics. These configurations ensure that the system remains efficient during training while performing well in terms of semantic alignment and spatial compatibility.

4.3 Semantic and image matching effects

To evaluate the performance of different cultural categories in terms of semantic consistency, accuracy of cultural background expression, style consistency, and cultural expression score, this experiment designs a series of tests covering four major cultural categories: history and traditional culture, modern and urban culture, nature and environmental culture, and art and creative expression. By collecting and analysing the participants' evaluation data on these cultural categories, the performance differences of different cultural categories on specific indicators and their potential reasons are revealed. Figure 3 shows a detailed analysis of the experimental results.

Figure 3 Scores of different rating dimensions in different cultural categories, (a) semantic consistency score (b) cultural background expression accuracy score (c) content matching rating (d) style consistency score (e) cultural expression rating (see online version for colours)



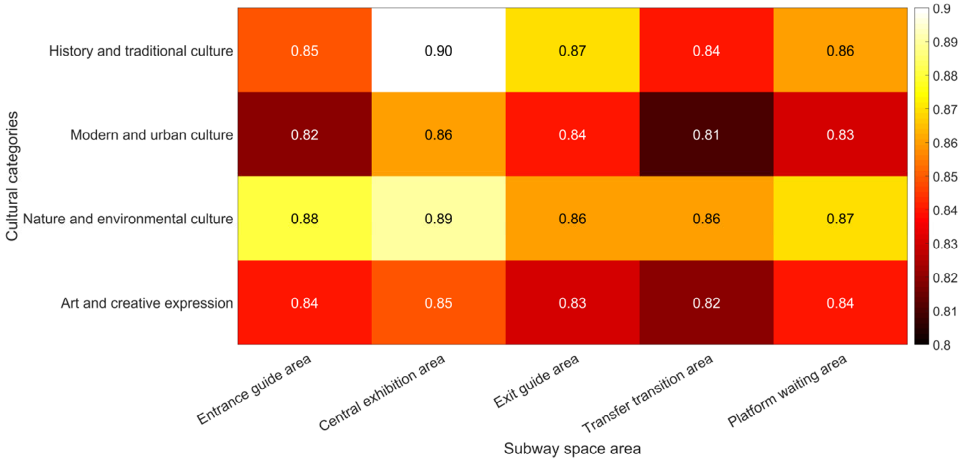
As can be seen from Figure 3, each cultural category shows high average scores in semantic consistency and cultural background expression accuracy, which are 0.9025 and 0.88, respectively, indicating that these cultural categories have strong coherence and accuracy in conveying their core meanings and cultural backgrounds. A comparative analysis is conducted across cultural categories to examine the influence of cultural differences on matching performance. The semantic consistency, background expression accuracy, style consistency, and cultural expression scores are analysed separately for each category, revealing that modern and urban culture achieves higher style consistency scores compared to history and traditional culture, while art and creative expression exhibits the highest cultural expression scores. These differences suggest that specific cultural attributes interact with the alignment mechanism in distinct ways, informing targeted optimisation strategies for each category. In terms of style consistency, the average score of all cultural categories is 0.885 points, which is related to the diversity of visual or language styles in different cultural categories, making it difficult to achieve a higher level of style consistency. In addition, the cultural expression score shows a similar trend, with a relatively balanced score across cultural categories and an average score of 0.86, reflecting the complexity of cultural expression in subjective perception. Overall, although there are slight differences in some indicators among different cultural categories, the overall performance is relatively stable, indicating that these cultural categories have strong consistency and effectiveness in conveying their core values and characteristics.

4.4 User visual hotspot coverage

To explore the changes in attention paid to different cultural themes in the multi-regional spatial distribution of urban subway public art spaces, this study focuses on five typical subway space areas: ‘entrance guidance area’, ‘central display area’, ‘exit guidance area’, ‘transfer transition area’, and ‘platform waiting area’, and sets four types of cultural content: ‘history and traditional culture’, ‘modern and urban culture’, ‘nature and environmental culture’, and ‘art and creative expression’. With the help of the user attention prediction network, the acceptance and attention intensity of various cultures in different spaces are quantitatively modelled and visualised, thereby revealing the coupling characteristics between cultural content and space type. Figure 4 shows the heat distribution of user attention scores for each cultural category in different subway space areas.

As shown in Figure 4, the attention of ‘history and traditional culture’ in the central display area reaches 0.90, which is the highest value of this cultural category in all spatial areas, indicating that it is more attractive in areas with strong spatial permeability and a large visual range. In the ‘nature and environmental culture’ category, the entrance guidance area and the central display area show high attention of 0.88 and 0.89, respectively, reflecting that users show a positive response to cultural expressions with soothing and ecological images when they initially enter the subway environment or wait for the train. ‘Art and creative expression’ scores relatively evenly in all spaces, with the central display area scoring slightly higher at 0.85, indicating that this type of cultural content has strong spatial adaptability and aesthetic universality. In contrast, ‘modern and urban culture’ has the lowest attention in the transfer transition area, at only 0.81, indicating that in a space dominated by functional access, abstract or modern symbolic cultural expressions fail to effectively arouse user attention. Overall, there is a significant interactive relationship between different cultural types and spatial functions, and specific cultural content is more likely to arouse user attention and cognitive investment in specific spaces.

Figure 4 User visual hotspot coverage of different cultural categories and hotspot areas (see online version for colours)



4.5 verification of the optimisation effect of the generative model based on user behaviour data

To further explore the impact of different cultural types on user emotional response and cognitive processing, this study designs a user evaluation experiment to evaluate users' subjective perception of four types of cultural content from four dimensions: emotional feedback, willingness to revisit, immersion, and cognitive load. The experiment collects scoring data in the form of a multidimensional scale to ensure that the indicators cover the main psychological and behavioural aspects of user experience. The performance differences of various cultures in user interaction are compared through structured analysis to reveal their potential design value and communication characteristics. The scoring results of various dimensions are summarised in Table 1.

Table 1 User sentiment and cognitive dimension rating performance

Cultural categories	*Emotional feedback score (out of 5 points)*	*Willingness to return visit score (out of 5 points)*	*Immersion score (out of 5 points)*	*Cognitive load score (out of 5, lower is better)*
History and traditional culture	4.3	4	3.8	2.5
Modern and urban culture	4.6	4.5	4.2	2
Nature and environmental culture	4.1	3.9	3.5	2.2
Art and creative expression	4.8	4.7	4.5	1.8

From the data results in Table 1, art and creative expression shows relatively superior user experience in all four dimensions. Its emotional feedback score is 4.8 points; the willingness to revisit is 4.7 points; the immersion score is 4.5 points; the cognitive load score is the lowest, only 1.8 points, showing the dual advantages of this cultural form in stimulating emotions and maintaining user concentration. The comprehensive score of modern and urban culture also ranks at the top, with emotional feedback and immersion scores of 4.6 and 4.2, respectively, indicating that this type has a high degree of reality relevance and situational appeal. In contrast, historical and traditional culture and natural and environmental culture are both lower than 4.5 points in emotional feedback and immersive experience, and their cognitive load scores are 2.5 and 2.2, respectively, suggesting that there is a certain degree of complexity in the way of information expression or content structure, which affects the cognitive processing efficiency of users. Overall, the degree of fit between the creative presentation of cultural content and the modern context plays a key role in improving users' positive emotions and cognitive fluency.

Table 2 Average ratings of different user groups on the five content perception dimensions

User groups	*Content understanding*	*Cultural resonance*	*Spatial correlation*	*Aesthetic perception*	*Information clarity*
18–30 years old	4.5	4.6	4.2	4.7	4.3
31–50 years old	4.3	4.4	4.0	4.5	4.1
51 years and above	4.0	4.1	3.8	4.2	3.9
Male	4.3	4.4	4.0	4.5	4.1
Female	4.4	4.6	4.1	4.6	4.2
Student	4.6	4.8	4.3	4.8	4.4
Office Workers	4.2	4.3	3.9	4.3	4.0
Cultural and creative professionals	4.5	4.7	4.4	4.9	4.5

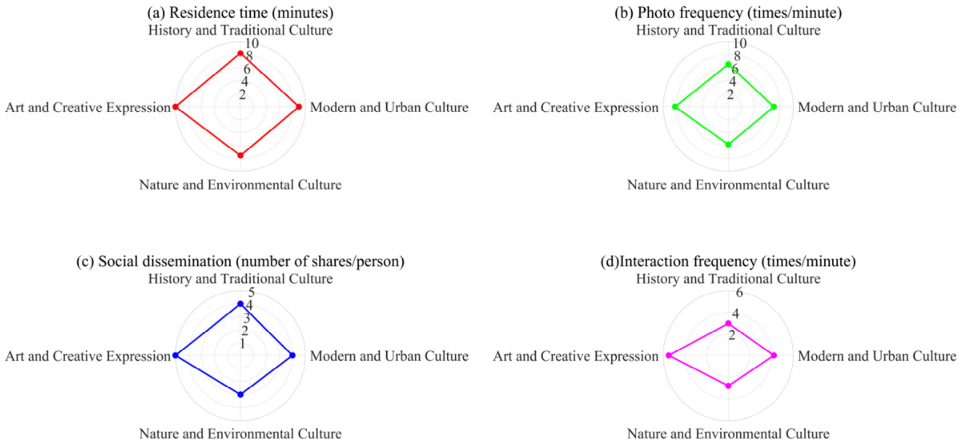
To further understand the acceptance and perception differences of users from different backgrounds towards generated content, this paper constructs a user segmentation evaluation system based on five dimensions, including content understanding, cultural resonance, spatial relevance, aesthetic perception, and information clarity, and conducts classification analysis based on the user's age, gender, and occupation. The average scores of the grouped users in the above dimensions are shown in Table 2.

As can be seen from Table 2, in the five scoring dimensions, the average score of the student group is always high, showing a high degree of adaptation to the content semantics, cultural connotations, and visual style. It reaches 4.6 points in the content understanding dimension and 4.8 points in cultural resonance and aesthetic perception, reflecting the high sensitivity of this group in context construction and picture-text association. Age level also has a significant impact on the scoring results. The group over 51 years old has the lowest scores in the spatial relevance and information clarity dimensions, 3.8 points and 3.9 points, respectively, indicating that the cognitive efficiency of middle-aged and elderly users in processing visual symbols and understanding the content transmission intention is relatively weak. Occupational identity has a significant pulling effect on aesthetic scores. Practitioners in the cultural and creative industry perform most prominently in the aesthetic perception dimension, with an average score of 4.9 points, which indirectly confirms their experience advantages in composition, colour, and cultural style recognition. Overall, the user's age, occupational background, and daily cultural contact frequency have a systematic impact on their subjective evaluation, which should be fully considered when optimising content generation strategies and oriented design.

To deeply understand the behaviour of users in multicultural fields, the system collects and quantifies user interaction data in different cultural theme scenes from four dimensions: residence time, photo frequency, social communication, and interaction frequency, to explore the potential relationship between cultural attributes and user engagement. This process constructs radar charts of four key indicators, horizontally comparing the differences in user behaviour in four cultural scenarios: 'history and traditional culture', 'modern and urban culture', 'nature and environmental culture', and 'art and creative expression', thereby revealing the effect pattern of cultural content in guiding user participation. Figure 5 shows the user behaviour radar chart formed by this analysis process.

In the comparison of the four dimensions, modern and urban culture performs at a higher level in terms of user retention time and photo frequency, which are 9 minutes and 7 times per minute, respectively. This shows that it has significant advantages in attracting user attention and stimulating recording behaviour. Nature and environment culture performs the worst in terms of social communication and interaction frequency, with a social communication degree of 3.8 times per person and an interaction frequency of 2.8 times per minute. This shows that this type of culture is not easy to stimulate users' willingness to spread and interact, which is related to its characteristics of relaxed scenes and soothing emotions. Although history and traditional culture have certain advantages in terms of photo frequency and retention time, they are insufficient in terms of interaction frequency and social communication, reflecting that they are affected by the traditional content expression form or limited interactive design. Art and creative expression performs best in all indicators, showing that they have a bright spot in stimulating user behaviour. Overall, different cultural themes have obvious differentiation characteristics in the dimension of behavioural triggering, reflecting the structural relationship between cultural content and the depth of user interaction.

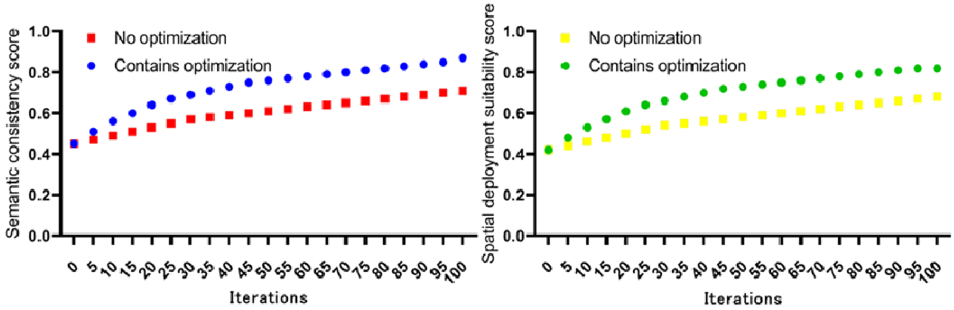
Figure 5 User experience dimension performance, (a) duration of stay by cultural category (b) frequency of taking photos by cultural category (c) social spread of different cultural categories (d) frequency of interaction by cultural category (see online version for colours)



4.6 Multi-objective optimisation effects

To evaluate the impact of the optimisation strategy on the model performance, this experiment designs two groups of comparative experiments: one group includes the optimisation strategy, and the another does not. Through the iterative training process, the experiment records the changing trends of the semantic consistency score and the spatial deployment suitability score. By analysing these two sets of data, this paper explores whether the optimisation strategy can significantly improve the performance of the model and further reveals the difference in its effects at different iteration stages, as shown in Figure 6.

Figure 6 Analysis of multi-objective optimisation effects, (a) comparison of semantic consistency score before and after optimisation (b) comparison of spatial deployment adaptability score before and after optimisation (see online version for colours)



(a) (b)

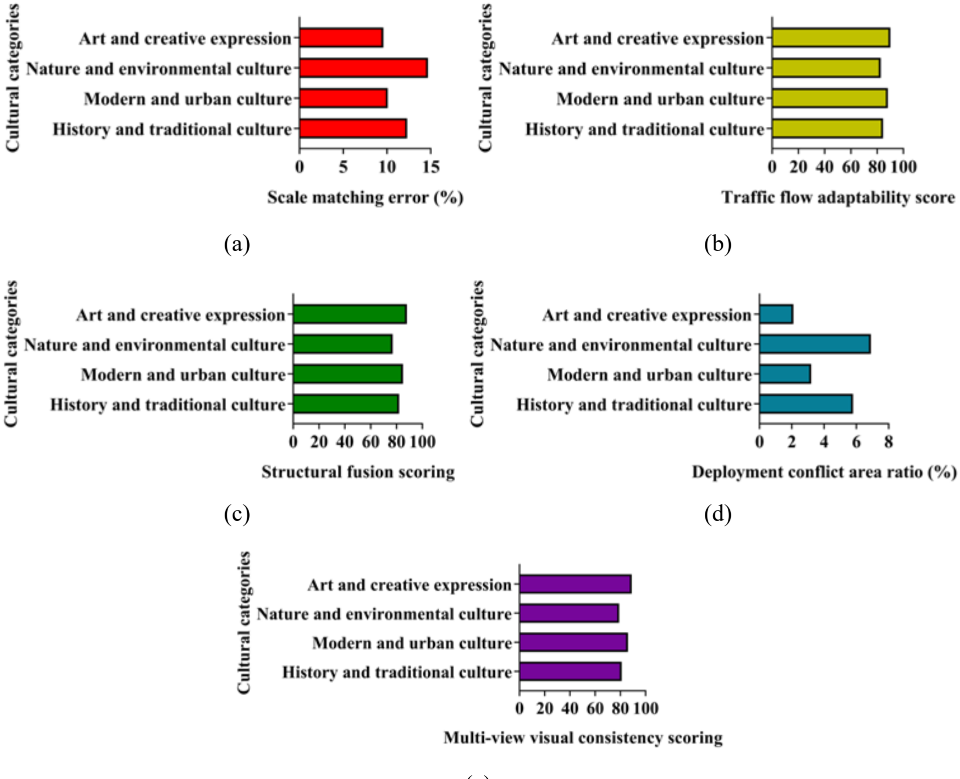
As can be seen from Figure 6, the model with the optimisation strategy performs better than the model without the optimisation strategy in both semantic consistency score and spatial deployment suitability score. Specifically, in terms of semantic consistency score, the model with the optimisation strategy is always higher than the model without the optimisation strategy throughout the iteration process. The gap gradually widens with the increase in the number of iterations, showing that the optimisation strategy effectively improves the model's semantic consistency. In terms of spatial deployment applicability scores, the model with optimisation strategy also shows higher scores and maintains steady growth in the later iterations, while the model without optimisation strategy grows slowly or even flattens. This difference is due to the fact that the optimisation strategy introduced a more effective parameter adjustment mechanism during the training process, thereby enhancing the model's adaptability and stability.

4.7 Spatial adaptability and structural fusion

In the process of digital modelling of cultural space, the study aims at the integration of multiple types of cultural elements in the urban public environment by constructing a multidimensional evaluation index system. From five dimensions, including scale matching error, traffic route adaptability, structural integration, proportion of deployment conflict areas, and multi-perspective consistency, the study conducts quantitative analysis on four types of cultural elements, namely 'art and creative expression', 'nature and environmental culture', 'modern and urban culture', and 'history and traditional culture'.

This process aims to explore the deployment potential and feasibility of different cultural elements in three-dimensional scenes, and form a criterion for judging their digital expression effects to support the subsequent optimisation of urban cultural visualisation design. Figure 7 shows the performance results of each cultural category under the five indicator dimensions.

Figure 7 Data analysis of spatial adaptability and structural integration indicators, (a) scale matching errors in different cultural categories (b) traffic flow adaptability scores for different cultural categories (c) structural fusion scores of different cultural categories (d) proportion of deployment conflict areas by cultural category (e) multi-view visual consistency scores for different cultural categories (see online version for colours)



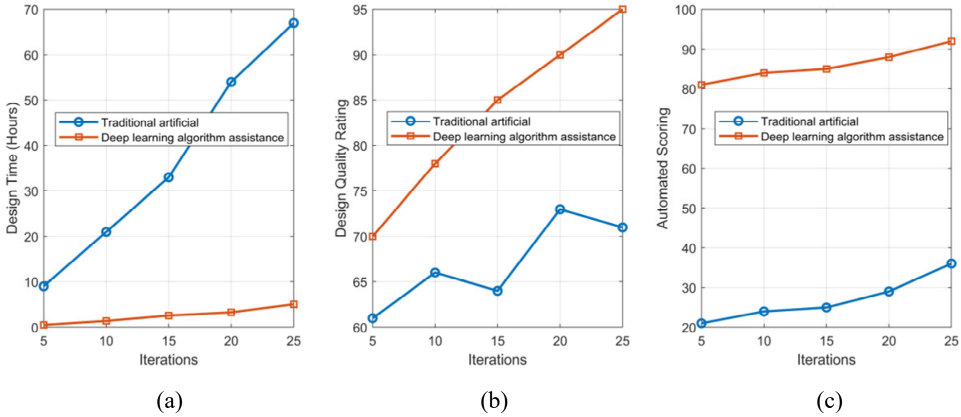
(a) (b) (c) (d) (e)

As can be seen from Figure 7, there are significant differences in the performance of different cultural categories in various indicators. In terms of scale matching error, natural and environmental culture shows the highest error value (14.7%), indicating that its adaptability in spatial scale is weak. Modern and urban culture shows a lower error value (10.1%), which is related to its simple design and high flexibility. In terms of traffic flow adaptability, historical and traditional culture receives a lower score (84.5 points), which is due to the complex design of traditional street layout and pedestrian flow patterns. In contrast, modern and urban culture scores higher (88.2 points), reflecting its good compatibility in responding to dynamic traffic needs. The structural integration score shows that the cultural category of art and creative expression scores the highest (88 points), indicating its strengths in innovation and diversity. The natural and environmental culture scores lower (77 points) because its emphasis on natural elements limits its integration with other elements. In terms of the proportion of deployment conflict areas, art and creative expression has the lowest rate (2.1%), reflecting its good compatibility with the existing urban structure; finally, in terms of multi-perspective visual consistency scores, the cultural category of art and creative expression once again performs well (89 points), reflecting its unity in visual aesthetics. In summary, different cultural categories have their own advantages and disadvantages in urban planning. Their differences in scale matching, traffic adaptability, structural integration, conflict area, and visual consistency provide important references for urban designers, which helps to select appropriate cultural category strategies according to specific needs in actual projects.

4.8 Efficiency comparison experiment of end-to-end generation pipeline

To evaluate the auxiliary effect of deep learning algorithms under traditional manual intervention, this experiment designs a comparative analysis of three key indicators: decision time, deep quality scoring, and automated scoring. The experiment gradually introduces the auxiliary function of deep learning algorithms in an iterative manner and compares it with traditional manual methods to explore its impact on task efficiency and quality. During the entire experiment, the changing trends of various indicators at different iteration times are recorded to reveal how deep learning algorithms optimise traditional manual processes. Next, the visualisation of the experimental results can be shown in Figure 8.

Figure 8 Comparison of the effects of traditional manual design and deep learning algorithm-assisted design, (a) design time comparison (b) design quality score comparison (c) automation level score comparison (see online version for colours)



As can be seen from Figure 8, when the traditional manual design and the deep learning algorithm-assisted design are compared with the same number of iterations, the design time of the latter is greatly reduced, showing that the deep learning algorithm can effectively improve the design efficiency. At the same time, the design quality score fluctuates under the traditional manual method, while the score assisted by the deep learning algorithm shows a clear upward trend, eventually reaching 95 points, showing the role of the algorithm in improving the task quality. In addition, the comparison of automation scores also shows a significant increase. The highest score of traditional manual design is 36, while the lowest score of deep learning algorithm-assisted design is 81, further verifying the contribution of deep learning algorithms in automation. Overall, the introduction of deep learning algorithms has accelerated the design process and significantly improved the quality and automation level, proving its effectiveness in optimising traditional manual processes.

In addition, this section also introduces three representative deep generation methods in the current urban art image generation task: StyleGAN2, Stable Diffusion v2, and DALL·E 2. They are compared with the multi-target diffusion generation framework that integrates CLIP adapter and user attention guidance in this paper, and the performance of the four in terms of design efficiency, image quality, semantic consistency, and spatial deployment adaptability is comprehensively evaluated. In the experimental setting, all models use the same cultural semantic label set and subway space scene images as input conditions. After generating batches of art images, the average generation time (in units of one image), image quality score (FID index), semantic consistency score (CLIP similarity score), and spatial adaptability score (fusion score based on the three-dimensional deployment evaluation mechanism) are recorded, respectively, as shown in Table 3:

Table 3 Comparison of three mainstream models and the proposed method on multi-dimensional performance indicators

Model name	*Average generation time (s)*	*FID score*	*CLIP semantic consistency*	*Spatial fit score*
StyleGAN2	3.8	29.2	0.782	71.5
Stable Diffusion v2	5.6	24.8	0.843	78.2
DALL·E 2	6.4	26.3	0.857	75.6
The method proposed in this paper	6.1	28.6	0.9025	86.8

Table 3 shows that in terms of average generation time, StyleGAN2 takes the lowest time of 3.8 seconds, while the method in this paper takes 6.1 seconds, slightly higher than the 5.6 seconds of Stable Diffusion v2. This difference is mainly due to the fact that the method in this paper introduces multi-objective optimisation and spatial adaptation calculation, which increases the computational complexity but improves the structural matching of the results. Although StyleGAN2 performs best in terms of FID (Fréchet Inception Distance) score of 29.2, its spatial adaptation score is 71.5, which is significantly lower than 86.8 of the method in this paper, indicating that while its generated content has high image quality, its spatial deployment capability is limited. In terms of semantic consistency, the proposed method scores 0.9025, which is significantly higher than DALL·E 2's 0.857. This is attributed to the fact that the semantic fine-tuning mechanism of CLIP-Adapter significantly enhances the accuracy of image-text alignment. The comprehensive index comparison results show that although the proposed method is slightly inferior in generation efficiency, it has comprehensive advantages in semantic expression accuracy and spatial structure fusion ability, and has stronger practical application potential.

5 Conclusions

Based on a multimodal semantic tagging system and a deep generation model, this study proposes an intelligent design method for urban subway public art that integrates cultural semantic alignment, user visual attention modelling, and spatial structure adaptation. The CLIP model is used to construct image-text semantic mapping, and the visual attention prediction network is used to simulate user behaviour to guide the artistic image generation process. The spatial adaptation analysis mechanism is then used to verify the feasibility of content deployment. The experimental results show that the semantic consistency score of this method reaches 0.9025, and the spatial deployment adaptation score is significantly improved after the introduction of the optimisation mechanism. The user attention of the ‘history and traditional culture’ category in the central display area is as high as 0.90, which verifies the response effect of the generated content in the visual focus area. In terms of automated scoring, deep learning-assisted design can reach up to 92 points, far exceeding the 36 points of traditional manual design, which significantly improves design efficiency and quality. However, when dealing with natural and environmental cultural labels, the spatial scale adaptation error of this method is relatively high (14.7%), reflecting that the adaptability problem in the semantic and spatial coupling process still needs to be optimised. Future research can further expand the adaptive generation mechanism in multicultural contexts, enhance the model’s ability to analyse and respond to complex spatial structures and multicultural connotations, and achieve a higher level of intelligent design of public art spaces.

Funding

Shaanxi Province ‘14th Five-Year Plan’ Education Science Planning 2022 Annual Project Practical Research on Talent Cultivation Mode of Private Colleges and Universities under the Perspective of Industry Clusters (SGH22Y1785).

Declarations

All authors declare that they have no conflicts of interest.

References

Benthien, C. (2021) ‘Public poetry: encountering the lyric in urban space’, *Internationale Zeitschrift für Kulturkomparatistik*, Vol. 2, pp.345–367, DOI: 10.25353/ubtr-izfk-271c-5517.

Campos, R. and Barbio, L. (2021) ‘Public strategies for the promotion of urban art: the Lisbon metropolitan area case’, *City & Community*, Vol. 20, No. 2, pp.121–140.

Castellano, G. and Vessio, G. (2021) ‘Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview’, *Neural Computing and Applications*, Vol. 33, No. 19, pp.12263–12282.

Chen, C. (2024) ‘Research on the emotionalized installation design of urban commercial spaces’, *Journal of China-ASEAN Studies*, Vol. 5, No. 1, pp.18–28.

Cheung, M., Smith, N. and Craven, O. (2021) ‘The impacts of public art on cities, places and people’s lives’, *The Journal of Arts Management, Law, and Society*, Vol. 52, No. 1, pp.37–50.

He, H. and Gyergyak, J. (2021) 'Enlightenment from street art activities in urban public space', *Pollack Periodica*, Vol. 16, No. 1, pp.169–175.

Kang, W., Gong, Q., Nakamura-Zimmerer, T. and Fahroo, F. (2021) 'Algorithms of data generation for deep learning and feedback design: a survey', *Physica D: Nonlinear Phenomena*, Vol. 425, No. 34, p.132955.

Kumar, S., Sastry, H.G., Marriboyina, V. et al. (2022) 'Semantic information extraction from multi-corpora using deep learning', *Computers, Materials & Continua*, Vol. 70, No. 3, pp.5021–5038

Lamas, D., Justo, A., Soilán, M., et al. (2024) 'Automated production of synthetic point clouds of truss bridges for semantic and instance segmentation using deep learning models', *Automation in Construction*, Vol. 158, p.105176, https://doi.org/10.1016/j.autcon.2023.105176.

Lee, J.M. (2022) 'Urban design in underground public spaces: lessons from Moscow Metro', *Journal of Asian Architecture and Building Engineering*, Vol. 21, No. 4, pp.1590–1605.

Lei, Y., Zhou, H., Xue, L., Yuan, L., Liu, Y., Wang, M. et al. (2024) 'Evaluating and comparing human perceptions of streets in two megacities by integrating street-view images, deep learning, and space syntax', *Buildings*, Vol. 14, No. 6, p.1847.

Lian, L. (2023) 'Systematic design and construction strategy of subway public art based on urban spirit', *Human Dynamics and Design for the Development of Contemporary Societies*, Vol. 81, No. 81, pp.18–28.

Matthews, T. and Gadaloff, S. (2022) 'Public art for placemaking and urban renewal: Insights from three regional Australian cities', *Cities*, Vol. 127, No. 3, p.103747.

Meng, Y. (2023) 'Analysis on the integral design of public art from the perspective of urban culture', *Journal of Humanities, Arts and Social Science*, Vol. 7, No. 2, pp.377–379.

Menghani, G. (2023) 'Efficient deep learning: a survey on making deep learning models smaller, faster, and better', *ACM Computing Surveys*, Vol. 55, No. 12, pp.1–37.

Milne, C. and Pojani, D. (2023) 'Public art in cities: what makes it engaging and interactive?', *Journal of Urban Design*, Vol. 28, No. 3, pp.296–315.

Pu, J. and Li, Y. (2023) 'Application of image style transfer based on normalized residual network in art design', *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 10, pp.38–45.

Ruthotto, L. and Haber, E. (2021) 'An introduction to deep generative modeling', *GAMM-Mitteilungen*, Vol. 44, No. 2, pp.1–26.

Sehar, U., Kanwal, S., Dashtipur, K. et al. (2021) 'Urdu sentiment analysis via multimodal data mining based on deep learning algorithms', *IEEE Access*, Vol. 9, pp.153072–153082, DOI: 10.1109/ACCESS.2021.3122025.

Shang, J. and Halabi, K.N.M. (2024) 'Urban subway space public art design exploration – taking Luoyang Metro Line 1 as an example', *International Academic Journal of Humanities and Social Sciences*, Vol. 4, No. 2, pp.11–11.

Shi, H., Chen, J., Feng, Z., Liu, T., Sun, D. and Zhou, X. (2025) 'Exploring the influence of environmental characteristics on emotional perceptions in metro station spaces', *Buildings*, Vol. 15, No. 3, p.310.

Šumak, B., Brdnik, S. and Pušnik, M. (2021) 'Sensors and artificial intelligence methods and algorithms for human-computer intelligent interaction: a systematic mapping study', *Sensors*, Vol. 22, No. 1, p.20.

Suzuki, M. and Matsuo, Y. (2022) 'A survey of multimodal deep generative models', *Advanced Robotics*, Vol. 36, Nos. 5–6, pp.261–278.

Wang, L. and Kim, J. (2022) 'Deep learning in computer real-time graphics and image using the visual effects of non-photorealistic rendering of ink painting', *Frontiers in Art Research*, Vol. 4, No. 15, pp.95–107.

Xinxin, L. and Hashim, A.M. (2024) 'Research on the application of interactive design in subway public art', *Asian Journal of Research in Education and Social Sciences*, Vol. 6, No. 3, pp.204–216.

Xu, Z. and Zhou, W. (2022) 'Metro public art from the perspective', *Human Factors in Architecture, Sustainable Urban Planning and Infrastructure*, Vol. 58, p.58, DOI: 10.54941/ahfe1002333.

Yang, J. (2023) 'Integrating operational efficiency, decorative design engineering and aesthetics in extended subway stations in China', *Operational Research in Engineering Sciences: Theory and Applications*, Vol. 6, No. 2, pp.294–314.

Zhang, Y. (2022) 'Modern art design system based on the deep learning algorithm', *Journal of Interconnection Networks*, Vol. 22, No. Supp05, 2147014.

Zhuoyu, G. (2023) 'The relationship between urban public space reconstruction and environmental art design', *Journal of Civil Engineering and Urban Planning*, Vol. 5, No. 9, pp.11–13.