ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Digital Cultural Tourism Visual Identity Design Based on AI Generative Models

Xiaorui Ye[1,*]

[1] Shaoxing Institute of Technology, Shaoxing City, Zhejiang Province 312000, China

SUMMARY: *Aiming at the problems of insufficient efficiency, unstable cultural expression and difficult style control in the visual image design of digital cultural tourism, this paper proposes a visual image design method for digital cultural tourism based on artificial intelligence generation model, and constructs a technical link of "demand analysis - data construction - multi-modal semantic fusion - generation control - feedback optimization". This study integrates deep learning, Transformer, multi-modal representation learning and diffusion generation technology, and constructs a dataset containing 12,000 cultural travel images, 8500 regional cultural texts, 3200 promotional texts and 2,100 groups of brand visual cases. On this basis, the generation of posters, logos, IP images and digital guide interfaces is completed. The experimental results show that the average FID of the model is 18.7, the SSIM is 0.842, the CLIP semantic similarity is 0.781, the cultural element fidelity is 88.4%, and the style matching degree is 86.9%. In the application case, the average generation time of four types of tasks is 6.33 s, and the average user satisfaction is 86.95%. The results show that this method can effectively improve the automation degree, cultural recognition degree and communication adaptability of digital cultural tourism visual design, and has strong theoretical value and application potential.*

KEYWORDS: *Visual image of digital cultural tourism; Artificial intelligence generative model; Multi-modal semantic fusion; Diffusion model*

1 Introduction

With the deep integration of digital technology, intelligent algorithms and the culture and tourism industry, the communication mode of culture and tourism is shifting from traditional static display to digital expression for multi-platform, multi-terminal and multi-scene. The visual image of cultural tourism brand is no longer a simple combination of logo, poster or publicity map, but an important medium for carrying regional culture identification, tourism context construction, user perception guidance and brand communication and transformation. Especially under the background of the continuous expansion of short video transmission, smart scenic spots, digital exhibition and online cultural tourism marketing, the frequency of visual content update is significantly accelerated, and cultural tourism brands have put forward higher requirements for design efficiency, style unity, scene adaptation and personalized expression. In contrast, traditional visual image design still relies heavily on designer experience and manual iteration, which has certain limitations in massive content generation, dynamic visual expansion, cross-media consistency control and rapid translation of cultural elements, and is prone to problems such as long production cycle, insufficient style

*yxr0702sx@163.com
https://doi.org/10.65102/is2026715
Published: 30 April, 2026

continuity and unstable expression of regional characteristics. Therefore, introducing the artificial intelligence generation model into the visual image design of digital cultural tourism is not only a technical update at the level of design tools, but also an important path to reshape the visual generation mechanism for the communication needs of digital cultural tourism.

From the research progress at home and abroad, the visual design of digital cultural tourism has gradually expanded from graphic design to a comprehensive system composed of image, video, interactive narrative and multimodal communication. Wang et al. (2024) studied the visual perception mode of tourism destination image based on inbound tourists' photos, and revealed the important role of tourism image content in the cognitive construction of destinations [5]. Tan et al. (2025) studied the relationship between multi-modal destination image and user participation, and explained that the image communication of cultural tourism has shown a trend of integration development of text, image and interactive feedback [6]. Wang (2025) studied the influence of anthropomorphic expression of destination advertising video on destination image supported by artificial intelligence technology, indicating that AI technology has begun to enter the core link of visual communication of cultural tourism brands [4]. In terms of generative artificial intelligence research, Zhu et al. (2025) proposed that tourist destination stereotypes would be significantly affected by GenAI-generated images, indicating that AI images are not only visual presentation tools, but also involved in tourism cognitive shaping [1]. Chung et al. (2025) studied the aesthetic saturation problem of generative AI images and pointed out that high-frequency repetitive AI visual content may weaken the freshness and dissemination effect of users [2]. Hou et al. (2025) studied the visual differences between AI-generated tourism photos and real tourism photos, human recognition and deep learning detection, reflecting that while AI images improve the realism, they also put forward new requirements for authenticity identification and content governance [3].

In the field of cultural heritage and digital cultural experience, the application of generative models further expands the technical boundaries of cultural tourism vision research. Ferracani et al. (2024) proposed to use AI visual illustrations to carry out personalized cultural heritage tourism narratives, proving that the generative model can enhance the visualization level of cultural knowledge transmission [7]. He et al. (2025) studied the process of collaborative construction of cultural heritage narratives by users and generative artificial intelligence, indicating that human-computer co-creation is becoming an important way of cultural content generation [8]. Fu et al. (2024) studied the display mechanism of GenAI co-creation achievements in cultural heritage communication and pointed out that it helped to improve audience participation and story experience [9]. Xu et al. (2025) proposed to introduce generative artificial intelligence into the narrative visualization process of interactive cultural experience, which provided new ideas for immersive visual design in digital cultural tourism scenes [10]. Cardarelli (2024) proposed the use of AI generation method to realize the digital reconstruction of archaeological objects [11], and Cardarelli (2025) further studied the application of one-step diffusion model in the generation of archaeological drawings [12], indicating that the diffusion model has strong potential in the visual reconstruction and specialized output of cultural objects. Li et al. (2025) studied the visual expression method of cultural heritage expert knowledge, which provided reference for cultural knowledge structuring and visual translation [13]. Kuang et al. (2025) proposed an automatic generation framework of historical building facades based on stable diffusion model [14]. Xiong and Wang (2025) studied the innovation path of traditional patterns by combining shape grammar and diffusion model [15]. Zhou et al. (2025) studied the method of fine-tuned diffusion model to generate a new design of intangible cultural heritage kite [16],

Hu et al. (2025) proposed an image-guided diffusion model for ancient mural restoration [17], Zou et al. (2025) studied a method of generating Chinese intangible cultural heritage images with structure perception and color perception [18]. Xu et al. (2025) systematically reviewed the development and application of generation technology in digital museums [19]. These results show that generative models, multi-modal representations and visual constraint mechanisms are driving cultural content expression from static display to intelligent generation.

However, there is still room for further research. On the one hand, the research on the visual image design of digital cultural tourism has not yet formed a unified framework with the extraction of regional cultural semantics, visual feature coding and coupling of brand communication needs as the core. On the other hand, although the existing generation research improves the image quality, there are still some deficiencies in the expression of regional characteristics, the consistency of series styles, the controllability of generated content and the evaluation mechanism of design results. Based on this, this paper intends to construct a framework for intelligent generation of visual images for digital cultural tourism scenes, and conducts research on cultural element extraction, visual semantic mapping, generative model-driven design and effect evaluation. In general, the technical route of "demand analysis, data construction, model design, visual generation, effect evaluation and application verification" is formed. The innovation of this paper is mainly reflected in three aspects. First, the semantic knowledge of digital cultural tourism is integrated with the generation model to enhance the cultural pertinence of visual output. The second is to construct an automatic generation method for cultural tourism visual image design to improve design efficiency and style stability. The third is to establish a multi-dimensional evaluation mechanism that takes into account image quality, cultural adaptability and communication effect. Therefore, this paper tries to establish a more interpretable and operable research path between the visual image design of digital cultural tourism and the integration research of computer technology.

2 The theoretical basis and Key Technology of digital cultural Tourism Visual Image Design

2.1 Dimensions of the visual image of digital cultural tourism

The visual image of digital cultural tourism is not a stack of single graphic elements, but a composite expression system composed of cultural semantics, visual coding and media adaptation. Its core dimensions mainly include regional cultural symbols, color system, graphic language, brand identity, IP image, interface visual style and media adaptation. Among them, regional cultural symbols assume the function of cultural recognition and scene reference, and semantic extraction can be completed through landmark buildings, historical patterns, folk customs and intangible cultural heritage elements. The color system is used to establish emotional perception and regional association, and form a unified visual tone through the hierarchical configuration of main color, auxiliary color and scene. Graphic language emphasizes shape structure, line style and layout order, which is the key link in the transformation of cultural elements into visual design objects. The brand identity and IP image assume the functions of memory reinforcement, personal communication and continuous operation. The visual style of the interface is oriented to digital guide, cultural tourism platform and interactive display scene, emphasizing graphic layout, dynamic feedback and consistency control. Media adaptation requires visual content to be able to achieve stable migration between posters, short video covers, mobile terminal interfaces and

immersive terminals. The above elements are not isolated from each other, but synergetic through the chain of "cultural semantics, visual generation and media output", forming the overall structure of the visual image of digital cultural tourism, as shown in Figure 1.

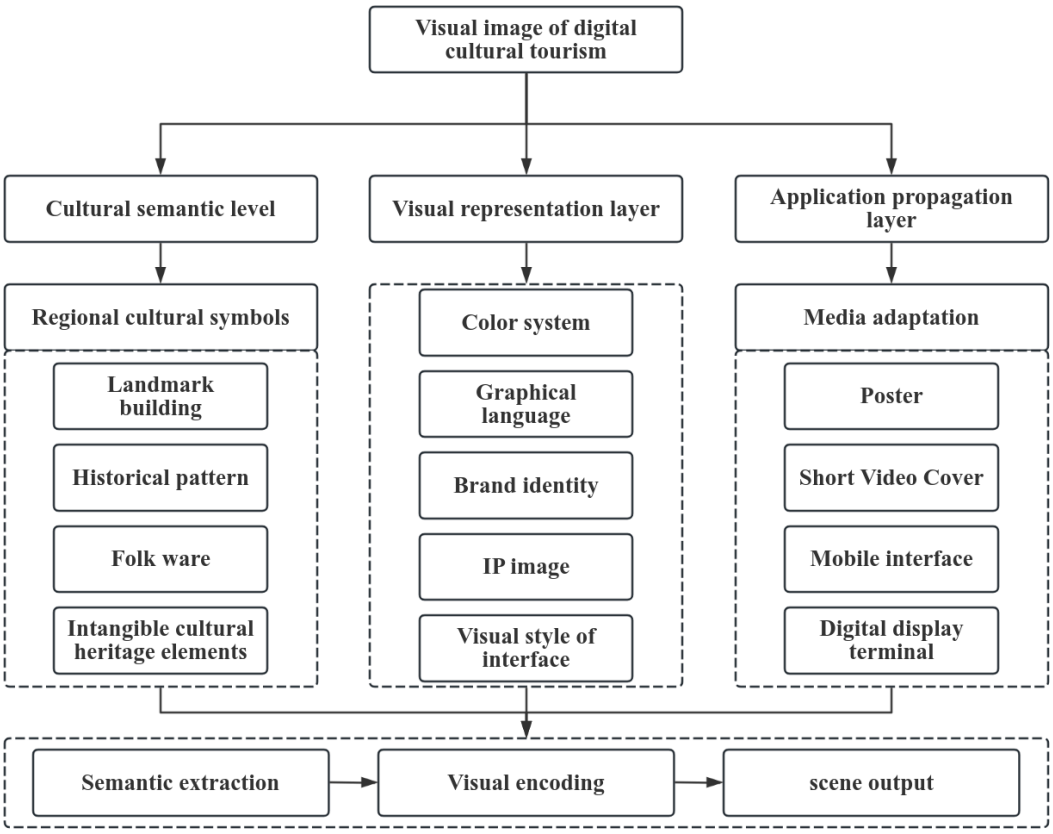

Figure 1: The visual image of digital cultural tourism constitutes a dimensional framework

2.2 The technical foundations of Generative Models for Artificial Intelligence

The technical foundation of artificial intelligence generation model is mainly built on the deep learning framework, and its core is to realize high-dimensional data feature extraction, semantic modeling and content generation through multi-layer nonlinear mapping. Convolutional neural network can extract local visual features such as edges, textures, shapes and spatial structures from images, so it is suitable for the recognition of landmark contours, pattern details and color distribution in cultural and travel images. Transformer relies on the self-attention mechanism to model long-distance dependencies, which can more effectively deal with the correlation between text description, visual semantics and cross-region image features, and provide global semantic support for visual generation in complex scenes. Multi-modal representation learning further realizes the unified embedding expression of text, image and cultural label, and forms a computable mapping between "regional culture semantics, visual design elements and generation target", which is an important prerequisite for the generation of digital cultural tourism visual image. In terms of the generation mechanism, the generative adversarial network improves the image realism through the game optimization of the generator and the discriminator, which is suitable for brand sketch generation, style transfer and creative scheme expansion. The diffusion model achieves high-quality image generation by gradually adding noise and inverse diffusion reconstruction,

which has better performance in detail fidelity, semantic consistency and style control, and is more suitable for cultural tourism posters, IP images and interface main visual design. The above technologies together constitute the computational basis for the intelligent generation of digital cultural tourism visual image, which makes the visual design change from experience-driven to data-driven and semantic-driven collaborative optimization.

2.3 Key Computer Technologies in Visual Generation of digital Cultural Tourism

The visual generation of digital cultural tourism is not a single image output process, but a computational chain composed of text semantic modeling, visual feature extraction, cross-modal alignment, generation control and result optimization. Firstly, in the text encoding stage, natural language cues such as scenic area names, regional cultural imagery, intangible cultural heritage elements, color descriptions and communication scenes need to be transformed into computable semantic vectors. Let the input text sequence be T={w1,w2,... wn}, the text representation can be obtained after word embedding and Transformer encoding:

$$z_t = \mathrm{Encoder}_t(T) \tag{1}$$

Among them, zt represents the global feature of cultural tourism semantics in the latent space, which is used to constrain the subsequent image generation direction. Different from traditional keyword matching, this process can preserve complex semantic relationships such as "ancient city texture", "landscape artistic conception", "national patterns" and "festival atmosphere".

In the stage of image feature extraction, convolutional neural networks or vision transformers are usually used to encode the sample images of cultural travel. Let the input image be I, and its visual features can be expressed as follows.

$$z_i = \mathrm{Encoder}_i(I) \tag{2}$$

Among them, zi contains information such as color distribution, texture structure, contour morphology and spatial layout. For the visual image of digital cultural tourism, this step is particularly important, because the outline of regional architecture, the rhythm of traditional patterns, and the composition method of tourism scenes often determine whether the design result has a clear place identification.

In order to achieve a consistent mapping from text description to visual content, cross-modal alignment is required. Generally, contrastive learning is used to make the matched text-image pairs closer and the mismatched samples farther apart in the feature space. The objective function can be written as follows.

$$\mathcal{L}_{\mathrm{align}} = -\log \frac{\exp(\,\mathrm{sim}(z_t, z_i)/\tau)}{\sum_{j=1}^{N} \exp(\,\mathrm{sim}(z_t, z_{i,j})/\tau)} \tag{3}$$

where sim(·) is the cosine similarity and τ is the temperature coefficient. This process can map "regional cultural semantics" and "image visual structure" into the same latent space, which provides a basis for subsequent generation control.

In the optimization stage of cue words, the initial cue words are often too broad, which is easy to cause the generation results to deviate from the design goal of cultural tourism. Therefore, it is necessary to jointly model the topic semantics, style constraints and scene requirements. Let the original prompt be p0, and the optimized prompt can be expressed as

follows.

$$p^* = p_0 + \alpha p_c + \beta p_s + \gamma p_m \tag{4}$$

Here, pc represents cultural semantic constraints, ps represents style constraints, pm represents medium adaptation constraints, and α,β,γ are weight parameters. In this way, compound requirements such as "Huizhou architectural style, green color, and mobile terminal propaganda posters" can be encoded into the generation process to improve the pertinency of the output content.

Style control is a key link in the visual generation of digital cultural tourism, whose goal is to ensure the unity of brand graphics, IP image, poster main vision and interface design in serial applications. If the feature matrix Fl is used to represent the image feature of the l-th layer, Gram matrix is commonly used to describe the style distribution:

$$G_l = F_l F_l^T \tag{5}$$

Style loss can be defined as follows.

$$\mathcal{L}_{style} = \sum_l \| G_l^{gen} - G_l^{ref} \|_F^2 \tag{6}$$

Here, G_l^{gen} is the generated image style matrix and G_l^{ref} is the reference style matrix. This method can make the generated results consistent in hue, brush stroke, pattern density and composition rhythm, which is suitable for the continuous output of digital cultural tourism brand visual system.

In the image generation stage, diffusion model has become an important technical path for high-quality visual generation. The forward process is to gradually add noise to the real image, and the backward process is to gradually denoise and restore the image under the constraint of the text condition. Let the noise state at step t be xt, then its inverse diffusion process can be written as follows.

$$p_\theta(x_{t-1}|x_t, z_t) \tag{7}$$

The training target is usually the prediction noise error:

$$\mathcal{L}_{diff} = \mathbb{E}_{x,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t, z_t)\|_2^2] \tag{8}$$

Compared with the traditional generative adversarial networks, the diffusion model is more suitable for the generation of cultural tourism posters, theme illustrations and interface main vision in terms of complex scene details, local texture quality and semantic consistency.

After the generation is completed, post-processing is also required to improve the usability of the image in the actual propagation. Post-processing usually includes super-resolution reconstruction, color correction, edge optimization, format adaptation, and artifact elimination. If the original generated result is denoted as Ig and the post-processing function is denoted as Φ(·), the final output can be expressed as:

$$I^* = \Phi(I_g) \tag{9}$$

In cultural and tourism applications, this step helps to make the generated images meet the resolution requirements of poster printing, mobile display, short video cover cropping, and

digital display. It can be seen that text encoding, image feature extraction, cross-modal alignment, prompt word optimization, style control, image generation and post-processing are not independent of each other, but together constitute the key computer technology system of intelligent generation of digital cultural tourism visual image.

2.4 The Advantages and Challenges of AI Generative Models Empowering Visual Design

The artificial intelligence generation model brings a method transition from "auxiliary production" to "intelligent generation" for the visual design of digital cultural tourism. Relying on deep learning, multi-modal alignment and diffusion generation mechanism, it can complete the output of a large number of visual solutions in a relatively short time, significantly improving the design efficiency. At the same time, the generative model can expand creative combinations through latent space sampling and conditional constraints, enhancing the exploration ability of graphics, color, and composition. In specific applications, the style transfer technology can realize the rapid conversion between traditional patterns, regional symbols and modern visual languages, and the personalized generation mechanism can output differentiated solutions according to user preferences, communication platforms and scene tasks, as shown in Table 1. However, there are still obvious challenges in the practical application of such models. On the one hand, cultural and tourism cultural elements have strong regional and context dependence, and if the training data or cue constraints are insufficient, it is easy to cause cultural expression distortion and symbol misuse. On the other hand, the model generation results may be affected by the distribution of training samples, resulting in aesthetic convergence, visual deviation and style homogeneity. In addition, the internal decision-making process of the diffusion model and the adversarial generation model is complex, and the generation path lacks full interpretability, which also affects the controllable correction and professional evaluation of the design results. Therefore, while enabling visual design, artificial intelligence generation models still need to combine cultural knowledge constraints, manual screening, and evaluation feedback mechanisms to achieve high-quality, credible, and usable output.

Table 1: Advantages and challenges of AI generative models enabling visual design

dimension	Main performance	Value of technology	Real-world challenges
Design efficiency	Quickly generate multi-version vision schemes	Shorten the design cycle and improve the iteration speed	It is easy to produce redundant schemes, and the screening cost increases
Creative development	Latent Space Composition and Diverse Sampling	Enhance the richness of ideas and the novelty of solutions	It may deviate from the semantics of cultural and tourism topics
Style transfer	Traditional elements are transformed into modern visual expressions	Enhance visual innovation and cross-style adaptation	Over-transfer of style can weaken cultural authenticity
Personalized generation	Customize the output for the user and scenario	Improve the matching degree of communication and interactive experience	The effect is unstable when the individual preference modeling is insufficient
Cultural expression	Blend symbol, color and narrative imagery	Supporting digital translation of regional culture	It is prone to symbol misreading and content distortion
Model mechanism	Automatically learn complex visual regularities	Improve generation quality and detail representation	Weak interpretability and difficult to accurately intervene

3 Visual image design method of digital cultural Tourism based on artificial intelligence generation model

3.1 Digital cultural tourism visual image design needs analysis

The visual image design requirements of digital cultural tourism are not a linear summary of single aesthetic needs, but a multi-dimensional task set formed by the joint action of cultural tourism brand communication objectives, regional cultural expression requirements, tourists 'aesthetic preference characteristics and digital media application scenarios. In order to improve the controllability and output adaptability of the subsequent generation model, it is necessary to transform the design requirements into a computable structured representation. In this paper, the requirements analysis process is divided into four steps: semantic acquisition, feature coding, weight evaluation and task mapping. The semantic acquisition is mainly for scenic area introduction text, urban culture and tourism publicity materials, user comments, image samples and platform communication data. Relying on natural language processing, image feature extraction and multi-modal embedding technologies, feature coding maps the abstract cultural appeal and communication goal into a vector space that can be used for generative control.

In the dimension of brand communication, the design requirements are mainly reflected in recognition, memory and communication consistency. Brand communication demand can be expressed as a vector:

$$\mathbf{b} = [b_1, b_2, b_3, b_4] \tag{10}$$

where b_1 represents the strength of brand recognition, b_2 represents the strength of visual memory, b_3 represents cross-platform consistency, and b_4 represents the tendency of communication transformation. This vector can be obtained by text topic modeling, social media visual content clustering, and brand keyword frequency statistics. The dimension of regional culture expression emphasizes the visual translation of local architecture, historical context, folk symbols, intangible cultural heritage patterns and color images. Let the set of cultural elements be C={c1,c2,... ,cn}, whose importance weight can be calculated by the attention mechanism:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \tag{11}$$

where ei represents the semantic contribution of the ith cultural element in the context of culture and tourism, and αi is used to guide the generation model to preferentially retain high-value cultural features.

The dimension of tourists' aesthetic preference mainly reflects the target audience's preference for color style, composition method, graphic complexity and emotional atmosphere. This part can combine user comment text, click behavior, stay time and interaction data to model. If the user preference feature vector is u and the design style feature vector is s, the matching degree of the two can be defined as follows.

$$\text{Score}(\mathbf{u}, \mathbf{s}) = \frac{\mathbf{u} \cdot \mathbf{s}}{\|\mathbf{u}\|\|\mathbf{s}\|} \tag{12}$$

The larger this index is, the more consistent the design style is with the target user's aesthetic preference. The application scenarios of digital media determine the output

specifications and interaction modes of visual content, and there are obvious differences in the requirements of resolution, layout density, information level and dynamic adaptation in different scenarios. A media scene can be represented as a set M={m1,m2,... ,mk}, such as mobile interfaces, short video covers, digital display screens, social media posters, etc., and build a "demand-scene" mapping function:

$$f:(\mathbf{b},\mathbf{c},\mathbf{u})\rightarrow M \tag{13}$$

Accordingly, the specific parameter configuration of the generation task is determined.

Considering the above factors, this paper defines the visual image design task of digital cultural tourism as a multi-objective optimization problem. Let the overall demand vector be:

$$\mathrm{D}=\mathrm{w_b b}+\mathrm{w_c c}+\mathrm{w_u u}+\mathrm{w_m m} \tag{14}$$

where c is the cultural expression feature vector, m is the media adaptation feature vector, wb,wc,wu,wm are the corresponding weights, and it satisfies:

$$\mathrm{w_b}+\mathrm{w_c}+\mathrm{w_u}+\mathrm{w_m}=1 \tag{15}$$

Through the structured requirements system, the design pre-analysis that originally relied on empirical judgment can be transformed into the conditional input of the generative model, which provides a clear computing basis for the subsequent extraction of cultural elements, visual semantic mapping and image generation control, and also makes the visual image design of digital cultural tourism further move from subjective creation to data-driven and intelligent collaboration.

3.2 Construction and preprocessing of Cultural Tourism Visual Semantic Dataset

In order to ensure that the artificial intelligence generation model can accurately learn the cultural semantics, style features and media expression rules in the visual image of digital cultural tourism, it is necessary to construct a multi-source visual semantic dataset for cultural tourism scenes. The dataset is not simply a collection of cultural travel pictures, but is jointly organized around three types of information of "cultural content-visual element-communication scene", forming a training basis that can support text-driven generation and multi-modal mapping. Specifically, the data sources mainly include cultural and tourism images, regional cultural texts, tourism publicity materials and brand visual cases, as shown in Figure 2.

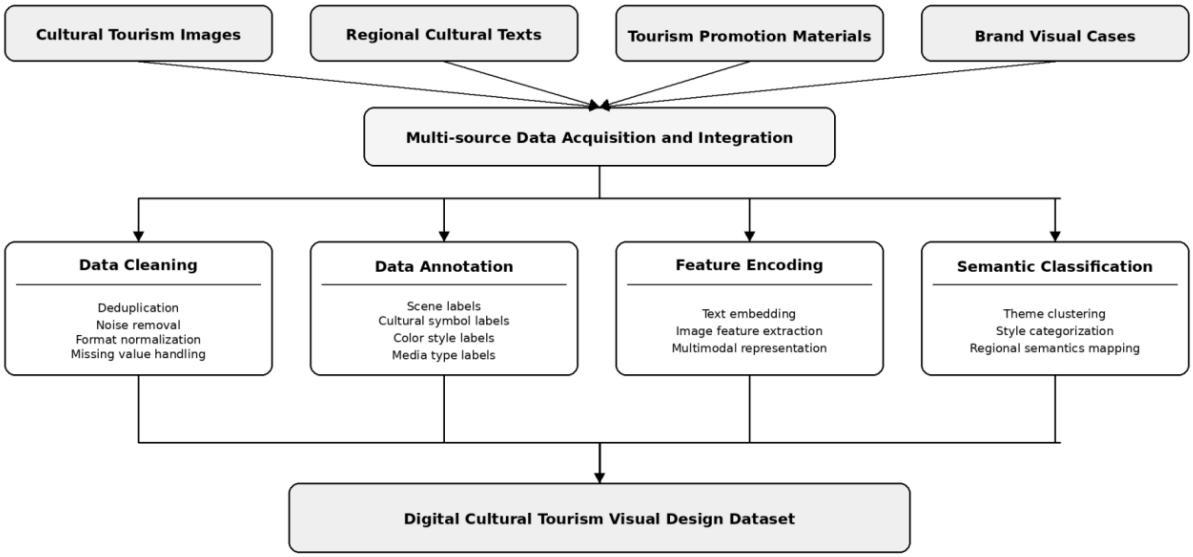

Figure 2: Workflow of Cultural Tourism Visual Semantic Dataset Construction and Preprocessing

Among them, cultural tourism images include scenic spot photos, publicity posters, guide interface screenshots, cultural tourism IP images and theme illustrations, which can be used to extract composition, color, texture and graphical features. Regional cultural texts include chorography, scenic spot introduction, intangible cultural heritage description, folk custom narrative and historical and cultural materials, which can be used to construct cultural semantic vocabulary and theme tags. Tourism promotional materials mainly include official tweets, short video copywriting, brand slogans and event planning texts, which are used to supplement communication objectives and audience oriented information. Brand visual cases include mature scenic spot logo system, urban cultural tourism main vision, digital display interface and cultural and creative packaging samples, which are used to learn serial style structure and design specifications.

In the data preprocessing stage, data cleaning, labeling, feature encoding and semantic classification need to be completed in turn. Data cleaning is mainly used to eliminate noisy samples, duplicate samples, and low-quality content. Let the original data set be D0={x1,x2,... xn}, after repeated detection, missing correction and format standardization, the cleaning results are obtained:

$$D_c = \phi(D_0) \tag{16}$$

Here, $\phi(\cdot)$ denotes the cleaning function. For image data, the resolution, color space and file format should be unified. For text data, it is necessary to perform word segmentation, stop words removal, error word correction and standardization of proprietary cultural vocabulary. This can reduce the noise interference and improve the stability of subsequent feature extraction.

In the data annotation stage, it is necessary to construct a multi-layer label system suitable for the visual design of digital cultural tourism. Image samples are usually annotated with cultural tourism scene categories, regional cultural symbols, color styles, composition patterns, and media types, while text samples are annotated with cultural themes, emotional tendencies, and communication purposes. Let's say that the label set of the third sample is:

$$Y_i = \{y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}\} \tag{17}$$

Then the supervised label space of the entire dataset can be expressed as follows.

$$\mathcal{Y} = \bigcup_{i=1}^{N} Y_i \tag{18}$$

This multi-label mechanism can simultaneously retain composite information such as "Huangshan cloud sea", "Hui-style architecture", "green color" and "mobile poster", so that the generative model can learn more fine-grained semantic constraints in the training phase.

In the feature encoding stage, text, image and other media attributes need to be mapped into a unified computational space. Let the text feature be Xt, the image feature be Xi, and the media attribute feature be Xm, then the sample comprehensive representation can be written as follows.

$$X = [X_t; X_i; X_m] \tag{19}$$

where $[\cdot]$ represents the feature concatenation operation. Text features are usually obtained by word embedding models or Transformer encoders, image features can be extracted by convolutional neural networks or vision Transformers, and media attributes are represented by one-hot encodings or low-dimensional embedding vectors. Through this multi-modal feature fusion method, abstract cultural semantics, specific visual information and application scene constraints can be integrated into a unified representation framework.

After the feature representation is completed, semantic classification is also required to form a category structure suitable for visual generation tasks. Let the classification function be $f(\cdot)$, then the semantic label prediction result can be expressed as follows.

$$\hat{y} = f(X) \tag{20}$$

The category assignment is of the form:

$$c^* = \arg\max_c P(c|X) \tag{21}$$

where $P(c \mid X)$ is the conditional probability of belonging to class c. Through theme clustering, style classification and regional semantic mapping, the cultural and tourism samples can be further organized into high-level categories such as "landscape class", "historical and ancient city class", "national culture class" and "non-corpse examination class", while retaining sub-level semantic information such as color style, graphic form and communication media. The final digital cultural tourism visual design dataset can not only support the extraction of cultural elements and visual semantic mapping, but also provide a stable data basis for subsequent diffusion models, multi-modal generation models and style control algorithms, so as to improve the accuracy, consistency and availability of digital cultural tourism visual image generation.

3.3 Feature Representation Method Based on Multi-modal Semantic Fusion

The key to the generation of digital cultural tourism visual image is not the accumulation of

single modal information, but the effective modeling of semantic association between text, image and cultural tags. For cultural and tourism scenes, text descriptions usually carry regional cultural imagery, communication themes and emotional appeals, image samples mainly reflect color distribution, composition structure and visual style, and cultural labels are used to describe landmark symbols, intangible cultural heritage elements, historical categories and media attributes. If the three types of information cannot be stably mapped in the same feature space, the generative model is prone to problems such as weakening of cultural semantics, drifting of style expression and homogenization of visual results. Therefore, this paper constructs a multi-modal semantic fusion model for visual design tasks, and realizes the joint representation of regional cultural imagery, color preference, symbol structure and style features through text coding, image coding, label embedding and cross-modal attention mechanism, as shown in Table 2.

Table 2: Main components of feature representations for multimodal semantic fusion

Modal type	Primary Data Sources	Extracting content	Calculation method	Output function
Text modality	Scenic spot introduction, intangible cultural heritage description, publicity copy, user comments	Regional cultural image, theme semantics, emotional tendency, communication appeal	Transformer Encoding, Word Embedding, Attention Modeling	Provides semantic constraints and topic orientation
Image modality	Scenic photos, visual posters, guide interface, brand cases	Color preference, texture details, composition structure, style characteristics	CNN/ViT Feature Extraction, Visual Encoding	Provides visual style and spatial information
Label modality	Regional categories, cultural symbols, media types, style labels	Cultural categories, symbolic structures, scene properties	Embedding mapping, label aggregation	Provide prior knowledge and class constraints
Fusion layer	Joint input of text, image and label	Correlating features across modalities	Feature mapping, Cross-modal Attention, Contrastive learning	Form a unified multimodal semantic representation
Output layer	Joint representation vector	Design condition vectors and style matching features	Similarity calculation, classification, or generation control	Support visual generation and effect optimization

Let the text input be T, the image input be I and the set of cultural labels be C={c1,c2,... ,ck}. Firstly, the Transformer text encoder is used to obtain the semantic representation of the text journey:

$$\mathbf{h}_t = \mathrm{Enc}_t(T) \tag{22}$$

Among them, ht∈Rdt represents the topic semantics, sentiment tendency and cultural

narrative features in the text. For image modalities, convolutional neural networks or vision transformers are used to extract visual features:

$$\mathbf{h}_i = \mathrm{Enc}_i(I) \tag{23}$$

Among them, hi∈Rdi mainly contains color statistics, texture details, contour structure and spatial layout information. Cultural labels are mapped to low-dimensional semantic vectors by embedding matrix:

$$\mathbf{h}_c = \frac{1}{k}\sum_{j=1}^{k} \mathrm{Emb}\,(c_j) \tag{24}$$

Among them, hc∈Rdc is used to represent prior knowledge such as geographical attributes, cultural types, symbol categories, and communication scenarios. In this way, the semantic elements such as "Huangshan Mountain cloud sea", "Dunhuang mural color system", "Huizhou style horse head wall" and "intangible cultural heritage patterns" can be transformed into learnable computational representations.

After the completion of the modal encoding, it is necessary to further establish the association mechanism between text, image and label. In this paper, cross-modal attention is used for weight allocation. Taking the guidance of text to image features as an example, the attention weight can be expressed as follows.

$$\alpha_{ti} = \mathrm{softmax}\left(\frac{(\mathbf{W}_q\mathbf{h}_t)(\mathbf{W}_k\mathbf{h}_i)^{\top}}{\sqrt{d}}\right) \tag{25}$$

Here, Wq and Wk are the mapping matrices between query and key, respectively, and dd is the scaling factor. Thus, the text-guided visual enhanced representation is obtained as follows.

$$\tilde{\mathbf{h}}_i = \alpha_{ti}\mathbf{W}_v\mathbf{h}_i \tag{26}$$

Similarly, the modulation weight of cultural labels on text and image can also be calculated, so that cultural prior can play a constraint role in feature fusion. This process is able to highlight visual areas and semantic units that are highly related to the topic, and reduce the interference of irrelevant background, redundant style and weakly related elements on the generation task.

In order to form a unified multi-modal representation, this paper uses a combination of feature concatenation and linear mapping to fuse three types of features into a joint representation vector for visual design:

$$\mathbf{z} = \phi\left(\mathbf{W}_t\mathbf{h}_t + \mathbf{W}_i\tilde{\mathbf{h}}_i + \mathbf{W}_c\mathbf{h}_c + \mathbf{b}\right) \tag{27}$$

where Wt,Wi,Wc are the mapping matrices of each mode, b is the bias term, and $\phi(\cdot)$ is the nonlinear activation function. The fused z simultaneously contains cultural theme, visual style, and application constraint information, which can be used as conditional input for subsequent prompt word enhancement, image generation control, and style consistency optimization. In order to enhance cross-modal consistency, contrastive learning objective function can be further introduced as follows.

$$\mathcal{L}_{con} = -\log \frac{\exp(\,\mathrm{sim}(\mathbf{h}_t, \mathbf{h}_i)/\tau)}{\sum_{j=1}^{N} \exp(\,\mathrm{sim}(\mathbf{h}_t, \mathbf{h}_{i,j})/\tau)} \tag{28}$$

where sim(·) is the cosine similarity and τ is the temperature parameter. The loss function can make the semantic matching text and image closer in the feature space, so as to improve the correspondence accuracy of "cultural description-visual output".

In practical design tasks, fusion representation not only serves for general image generation, but also undertakes the function of style recognition and design constraint modeling. If the design attribute of comprehensive color preference, symbol structure and composition style is denoted as s, the visual design suitability can be defined as follows.

$$\mathrm{Match}(\mathbf{z}, \mathbf{s}) = \frac{\mathbf{z} \cdot \mathbf{s}}{\|\mathbf{z}\|\|\mathbf{s}\|} \tag{29}$$

The larger the value, the more consistent the current multimodal feature representation is with the target design requirements. Based on the above mechanism, the multi-modal feature representation model constructed in this paper can establish a stable mapping between regional culture semantics and visual expression, and provide a unified computing basis for the generation of digital cultural tourism brand identity, main visual poster, IP image and interface style.

3.4 Generative Model-based mechanisms for Generating Visual images

In order to realize the automatic design of the visual image of digital cultural tourism, this paper constructs a visual generation framework of "semantic input-conditional encoder-generative inference-consistent optimization-result output", which takes text prompts, cultural labels, reference images and style constraints as the generation conditions to drive the generation of posters, brand logos, IP images and interface main vision. Specifically, we first encode the scene description text T, the cultural semantic label C and the reference visual sample Ir into a joint condition vector:

$$\mathrm{z} = \mathrm{Fuse}(\mathrm{Enc_t}(\mathrm{T}), \mathrm{Enc_c}(\mathrm{C}), \mathrm{Enc_i}(\mathrm{I_r})) \tag{30}$$

Fuse(·) represents the multi-modal fusion function, and z is used to uniformly represent regional cultural imagery, color style, symbol characteristics and media output requirements. The condition vector can be used as the semantic guidance input of the diffusion model, and can also be used as the control variable of the GAN generator, so as to enhance the matching degree between the generated content and the cultural tourism topic.

In the diffusion generation path, the model completes image sampling by forward noise addition and reverse denoising. Let the true image be x0 and the noise state be xt at step t, then the inverse diffusion process can be expressed as follows.

$$\mathrm{p_\theta}(\mathrm{x_{t-1}}|\mathrm{x_t}, \mathrm{z}) \tag{31}$$

The training objective is to minimize the noise prediction error:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\mathrm{x},\epsilon,\mathrm{t}}[\|\epsilon - \epsilon_\theta(\mathrm{x_t}, \mathrm{t}, \mathrm{z})\|_2^2] \tag{32}$$

The proposed mechanism can gradually recover high-quality images under semantic constraints, which is suitable for the generation of high-resolution visual content such as

poster main vision, scene illustration and interface background. For identity and IP images with clearer structure requirements, conditional GAN can be used to achieve rapid generation, and its basic form is as follows:

$$G: \{z, n\} \rightarrow \hat{x}, \quad D(x, z) \rightarrow [0,1] \tag{33}$$

where n is random noise, $\hat{x}$ is the generated result, and the confrontation goal between the generator and the discriminator is as follows.

$$\min_G \max_D \ \mathbb{E}_{x \sim p_{data}}[\log D(x, z)] + \mathbb{E}_{n \sim p(n)}[\log(1 - D(G(n, z), z))] \tag{34}$$

Compared with diffusion models, GAN is more suitable for symbolic visual design tasks in terms of graph contours, local morphology, and fast iteration.

In order to improve the controllability and consistency of the generated results, we further introduce prompt engineering and constrained optimization mechanism. The original cue word p0 often can only express coarse-grained semantics, so the cue is enhanced by cultural constraints, style constraints and medium constraints:

$$p^* = p_0 + \alpha p_c + \beta p_s + \gamma p_m \tag{35}$$

where pc represents regional culture constraints, ps represents color and style constraints, pm represents media scene constraints, and α,β,γ are control weights. At the same time, in order to ensure the uniformity of serial visual output in brand tonality, the consistency loss function can be defined as follows.

$$\mathcal{L}_{con} = \lambda_1 \| f_{sem}(\hat{x}) - f_{sem}(x_{ref}) \|_2^2 + \lambda_2 \| f_{sty}(\hat{x}) - f_{sty}(x_{ref}) \|_2^2 \tag{36}$$

f_{sem} and f_{sty} represent semantic feature and style feature extraction functions, respectively, and xref is the reference standard sample. Through this mechanism, the model can maintain accurate cultural expression, coordinated color system and stable style characteristics when generating multiple types of visual objects. Finally, the digital cultural tourism visual image output by the system not only has strong automatic generation ability, but also can meet the technical requirements of controllability, continuity and scene adaptation in cultural tourism brand communication.

3.5 Visual generation result optimization and design feedback mechanism

In order to improve the usability, stability and communication adaptability of the digital cultural tourism visual image generation results, this paper introduces the visual result optimization and design feedback mechanism after the generation model output, and iteratively modifies the generation results through three types of information: image quality evaluation, semantic consistency evaluation and user preference feedback, forming a closed-loop design process of "generation-assessment-correction". This is shown in Figure 3. The core idea of this mechanism is to transform the visual generation process from one-time output into a sustainable optimization calculation process, so that the model can not only generate images, but also automatically adjust the cue words, constraints and parameter configurations according to the evaluation results, so as to gradually approach the optimal design results.

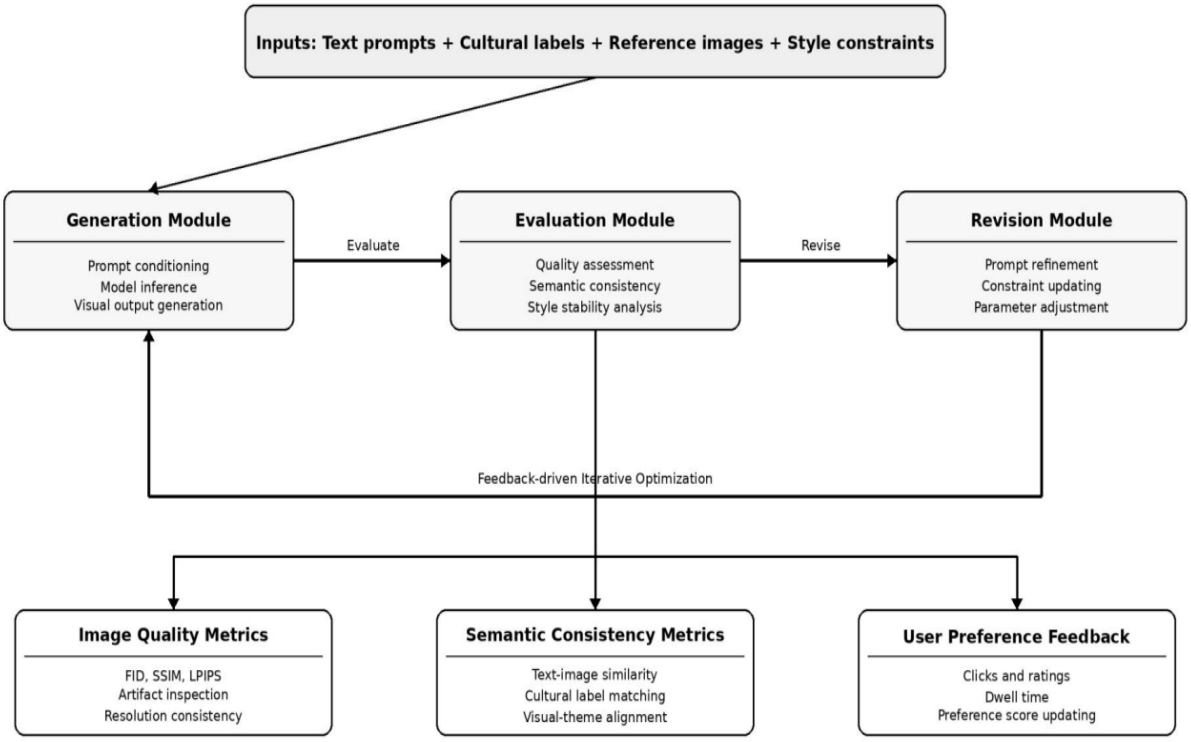

Figure 3:Closed-loop Workflow of Visual Generation Optimization and Design Feedback

At the image quality assessment level, it is necessary to focus on the clarity, structural integrity, detail fidelity and artifact control effect of the generated image. Let the generated image be $\hat{I}$ and the reference image be Ir, then the image quality loss can be expressed as follows.

$$L_q = \alpha_1 L_{fid} + \alpha_2 L_{ssim} + \alpha_3 L_{lpips} \tag{37}$$

L_{fid} is used to measure the difference between the generated distribution and the real image distribution, L_{ssim} reflects the structural similarity, and L_{lpips} characterizes the visual difference at the perception level. Through this evaluation module, problems such as blurred edges, texture breaks, color contamination and local artifacts in the main visual image can be identified, which provides a basis for subsequent correction.

At the level of semantic consistency evaluation, it is necessary to judge whether the generated results accurately express the theme of culture and tourism, regional cultural imagery and design task objectives. Let the text cue code be zt and the generated image code be zi, then the semantic consistency score can be defined as follows.

$$S_{sem} = \frac{z_t \cdot z_i}{\|z_t\| \|z_i\|} \tag{38}$$

The corresponding semantic loss can be written as follows.

$$L_s = 1 - S_{sem} \tag{39}$$

Although the generated images have high visual quality, but fail to effectively reflect the core semantics such as "landscape artistic conception", "ancient city texture", "national patterns" or "intangible cultural heritage color", the index can timely reflect the semantic drift

problem. At the same time, the cultural label matching test can also be superimposed to ensure that the generated results have a clear direction in regional cultural expression.

At the user preference feedback level, the system further judges whether the generated content meets the aesthetic preferences and communication needs of the target audience by collecting click through rate, stay time, scoring results and interactive behavior. Suppose the user feedback vector is u and the generated feature is $f(\hat{I})$, then the preference adaptation score can be expressed as follows.

$$S_u = \frac{u \cdot f(\hat{I})}{\|u\|\|f(\hat{I})\|} \tag{40}$$

The corresponding preference loss is as follows.

$$L_u = 1 - S_u \tag{41}$$

This process enables the system to examine the design effect from the "model perspective" to the "user perspective", avoiding only relying on the algorithm indicators and ignoring the real communication performance.

Based on the three evaluation results, the overall optimization objective is defined as follows.

$$L = \lambda_1 L_q + \lambda_2 L_s + \lambda_3 L_u \tag{42}$$

Here,λ1,λ2,λ3 are three types of evaluation weights: quality, semantics, and preference. According to the change of the loss function, the system can dynamically update the cue words, style weights and control parameters. If we remember that the prompt word or control parameter in round t is p(t), the update form can be written as follows.

$$p^{(t+1)} = p^{(t)} - \eta \nabla L \tag{43}$$

Here, η is the learning step size. In this way, the generated results can be continuously corrected in multiple rounds of feedback, and finally form a visual output that balances image quality, cultural accuracy, and user acceptance. Therefore, visual generation is no longer a one-way image synthesis process, but a closed-loop optimization system that integrates model calculation, evaluation analysis and design modification.

4 Experiment design and result analysis

4.1 Experimental environment and parameter setting

In order to verify the effectiveness of the digital cultural tourism visual image generation method constructed in this paper, the model training, generation reasoning and result evaluation are completed in the unified computing environment. The experimental platform uses Python 3.10 and PyTorch 2.1 deep learning framework, and the tool libraries such as CUDA 12.1, OpenCV, Transformers and Diffusers are used to realize multi-modal feature coding, diffusion generation and result post-processing. The hardware environment is configured with Intel Xeon Silver 4314 CPU, NVIDIA RTX 4090 24GB GPU, 128GB memory, and 2TB SSD storage to ensure stable operation of high-resolution image training and batch generation tasks. The experimental data set contains 12000 cultural images, 8500 regional cultural texts, 3200 tourism publicity texts and 2100 groups of brand visual cases,

which are divided into training set, validation set and test set according to 8∶1∶1. In the model training phase, the image input resolution is set to 512×512, the batch size is set to 16, the training round Epoch is set to 120, the initial learning rate is 2×10^{-4}, AdamW is used as the optimizer, and the weight decay coefficient is set to 1×10^{-5}. The text encoding dimension is set to 768, the multimodal fusion dimension is set to 1024, and the cue word guidance weight is set to 7.5. The evaluation process is carried out according to "generation output - quality evaluation - semantic consistency check - user preference feedback - result statistical analysis", and FID, SSIM, CLIP similarity and user score are used to systematically verify the performance of the model, so as to ensure that the experimental process has strong computer technical standardization and reproducibility.

4.2 Analysis of Visual Image generation effect of digital cultural tourism

From the test results, the model in this paper shows good comprehensive stability in the visual image generation task of digital cultural tourism, and achieves better results in image quality, visual consistency, cultural element fidelity and style matching. Based on the statistics of 300 groups of samples in the test set, the average FID of the generated results is 18.7, the SSIM is 0.842, the CLIP semantic similarity is 0.781, the style matching degree is 86.9%, and the cultural element fidelity is 88.4%. Among them, the poster main vision generation task is the most stable in overall clarity and hierarchical expression, the IP image generation performs better in contour integrity and color coordination, and the interface main vision has higher adaptability in information carrying and background style unity.

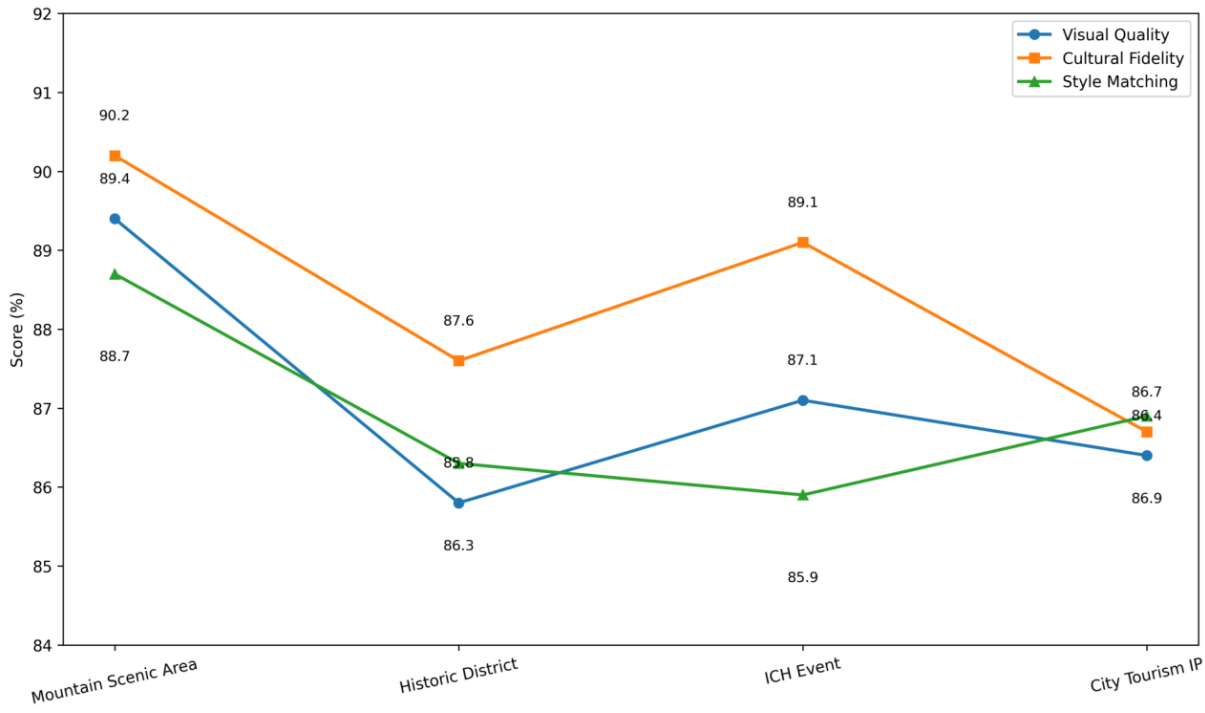

Figure 4: Comparison of Generation Performance Across Cultural Tourism Scenarios

Further divided by scene, there are certain differences in the results of visual quality, cultural fidelity and style matching of the four samples of landscape scenic spots, historical districts, intangible cultural heritage activities and urban cultural tourism IP. As shown in Figure 4, the three indicators of the samples of landscape scenic spots reach 89.4%, 90.2% and 88.7% respectively, and the overall performance is the best. The corresponding samples

of historical districts were 85.8%, 87.6% and 86.3%. The samples of intangible cultural heritage activities were 87.1%, 89.1% and 85.9%; The urban cultural tourism IP samples are 86.4%, 86.7% and 86.9%. This indicates that the model has a stronger reconstruction ability for visual tasks with clear geographic landscape boundaries and typical cultural symbols, while in IP design tasks with a high degree of abstraction, local semantic details still have some fluctuations.

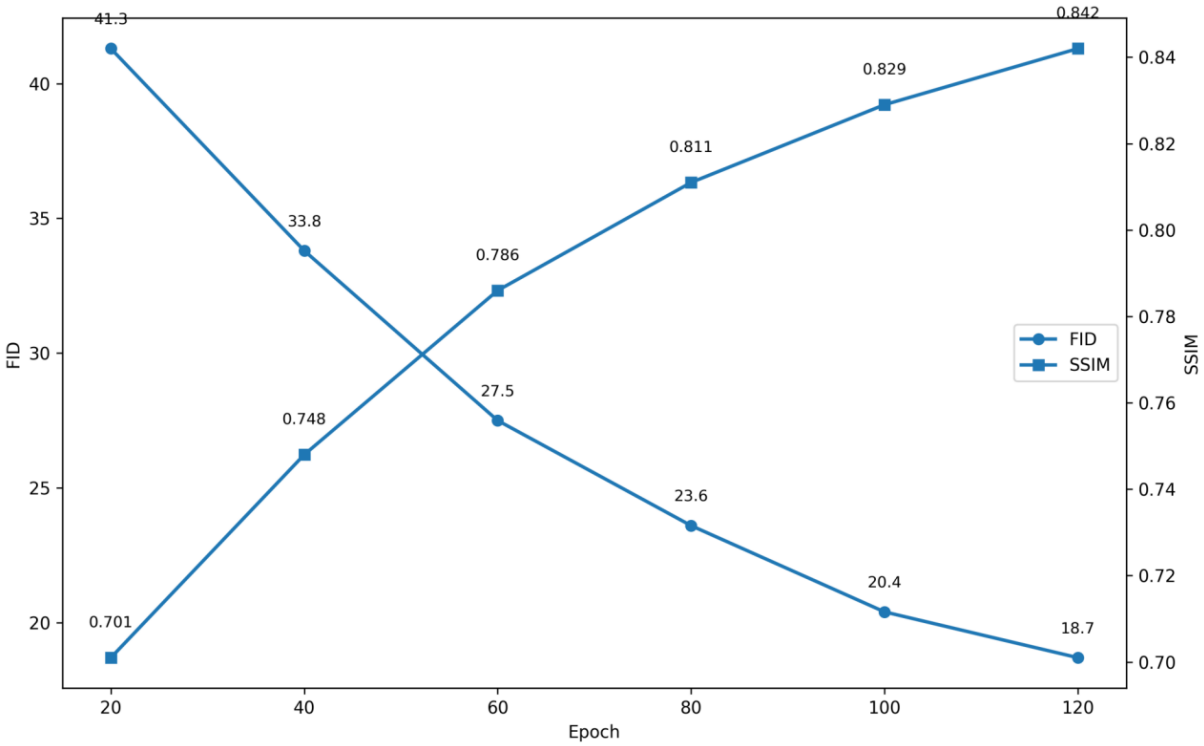

Figure 5: Training Performance Trends of the Digital Cultural Tourism Visual Generation Model

From the training process, the performance of the model gradually improves with iterations, as shown in Figure 5. When the training rounds increase from 20 to 120, the FID decreases from 41.3 to 18.7, which indicates that the difference between the generated image and the real sample distribution continues to decrease. At the same time, SSIM increases from 0.701 to 0.842, indicating that the structural integrity and visual stability of the image are continuously enhanced. Especially after 80 rounds, the decline rate of FID tends to be flat, while SSIM continues to improve slowly, indicating that the model gradually shifts from overall structure learning to detail optimization in the middle and late training stages. Combined with qualitative observation, it can be seen that the generated results can better retain the outline of ancient buildings, the traditional pattern structure, the regional color combination and the main visual atmosphere of the scenic spot, and reflect strong cultural recognition and scene adaptability in the digital poster, guide interface and cultural tourism brand image design.

4.3 Comparative experiments with different Generative models

In order to verify the advantages of the proposed method, diffusion model, conditional GAN, VAE and style transfer baseline are selected for comparative experiments, which are evaluated from four dimensions of generation accuracy, stability, detail performance and design

adaptability.

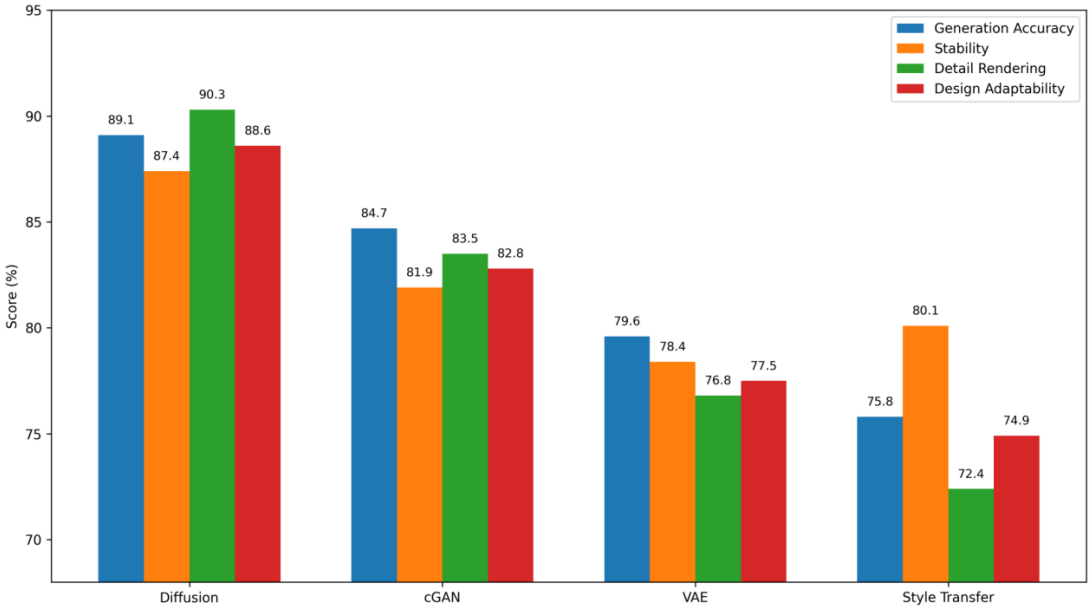

Figure 6: Performance Comparison of Different Generative Models for Digital Cultural Tourism Visual Design

The results show that the diffusion model has the best overall performance, as shown in Figure 6, with its four indicators reaching 89.1%, 87.4%, 90.3% and 88.6%, respectively, which are significantly better than the conditional GAN's 84.7%, 81.9%, 83.5% and 82.8%. The stability of VAE is acceptable, but the detail reconstruction ability is weak, and the detail performance is only 76.8%. The style transfer method has certain advantages in local style continuation, but limited by the original sample structure, the design adaptability is only 74.9%.

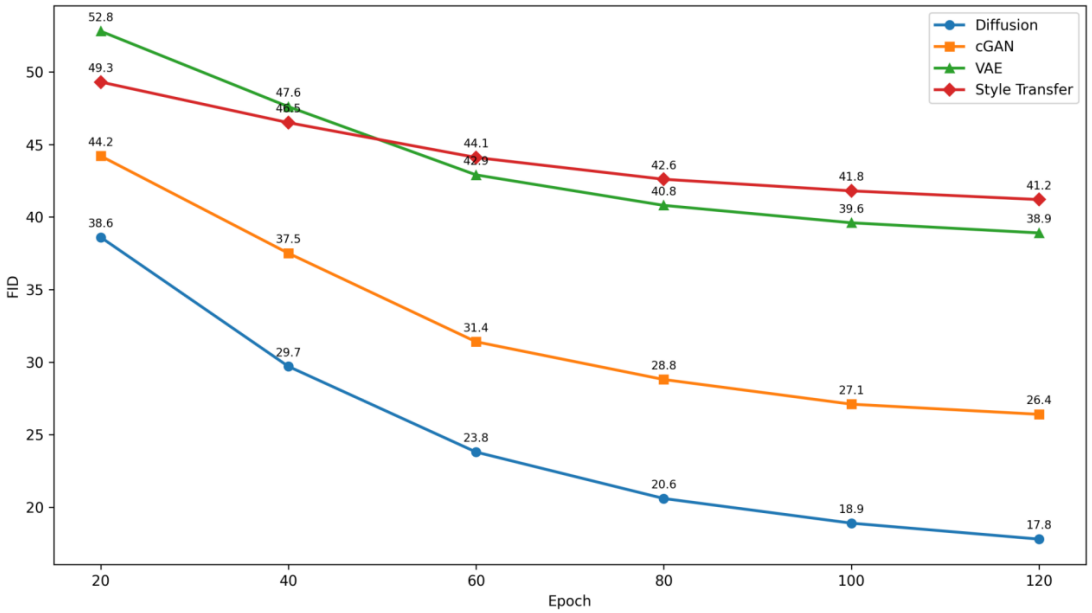

Figure 7: Training FID Trends of Different Generative Models for Digital Cultural Tourism Visual Design

From the perspective of the training convergence process, as shown in Figure 7, the FID of the diffusion model decreases from 38.6 to 17.8, which is the largest decline, indicating that it has more advantages in image distribution fitting and visual authenticity restoration. The final FID of conditional GAN is 26.4, which has a fast convergence speed, but the fluctuation is obvious in the later stage. The FID of VAE and style transfer model stays at 38.9 and 41.2, respectively, indicating their limited expressive ability in complex cultural tourism visual generation tasks. On the whole, the diffusion model is more suitable for the visual image design task of digital cultural tourism in terms of high-precision generation, cultural detail preservation and multi-scene design adaptation, while GAN is more suitable for fast iteration and symbolic graphics generation with clear contours.

4.4 Analysis of the Role of Multi-modal Semantic Fusion Mechanism

In order to verify the actual contribution of the multimodal semantic fusion mechanism to the visual image generation of digital cultural tourism, this paper designs an ablation experiment to remove the text semantics, image features and cultural label modules respectively, and compares them with the full model. The experimental results show that multi-modal fusion has a significant effect on improving the quality of generation, the accuracy of cultural expression and the stability of style.

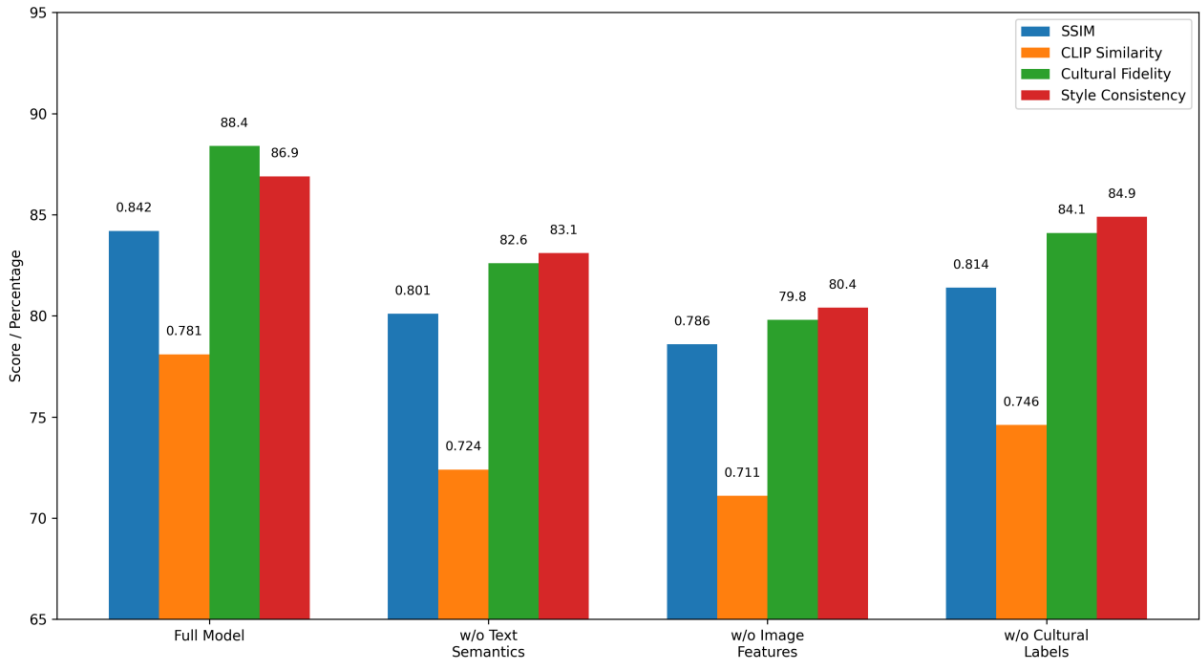

Figure 8: Ablation Performance Comparison of Multimodal Semantic Fusion Components

As shown in Figure 8, the complete model outperforms other ablation Settings in all indicators, with its SSIM reaching 0.842, CLIP semantic similarity 0.781, cultural element fidelity 88.4%, and style consistency 86.9%. When the text semantic module is removed, the model can still maintain a certain visual structure, but the cultural theme expression is weakened, the SSIM is reduced to 0.801, the CLIP similarity is reduced to 0.724, and the cultural element fidelity is reduced to 82.6%. When the image feature module is removed, the model has the most obvious decline in color restoration, texture details and composition stability, with the SSIM of only 0.786 and the fidelity of cultural elements reduced to 79.8%, indicating that image modality has a fundamental support role for visual detail and structure reconstruction. After removing the cultural label module, the degradation of the overall visual

quality of the model is relatively small, but the cultural scene recognition and style pertinence are weakened, and the cultural element fidelity and style consistency are reduced to 84.1% and 84.9%, respectively.

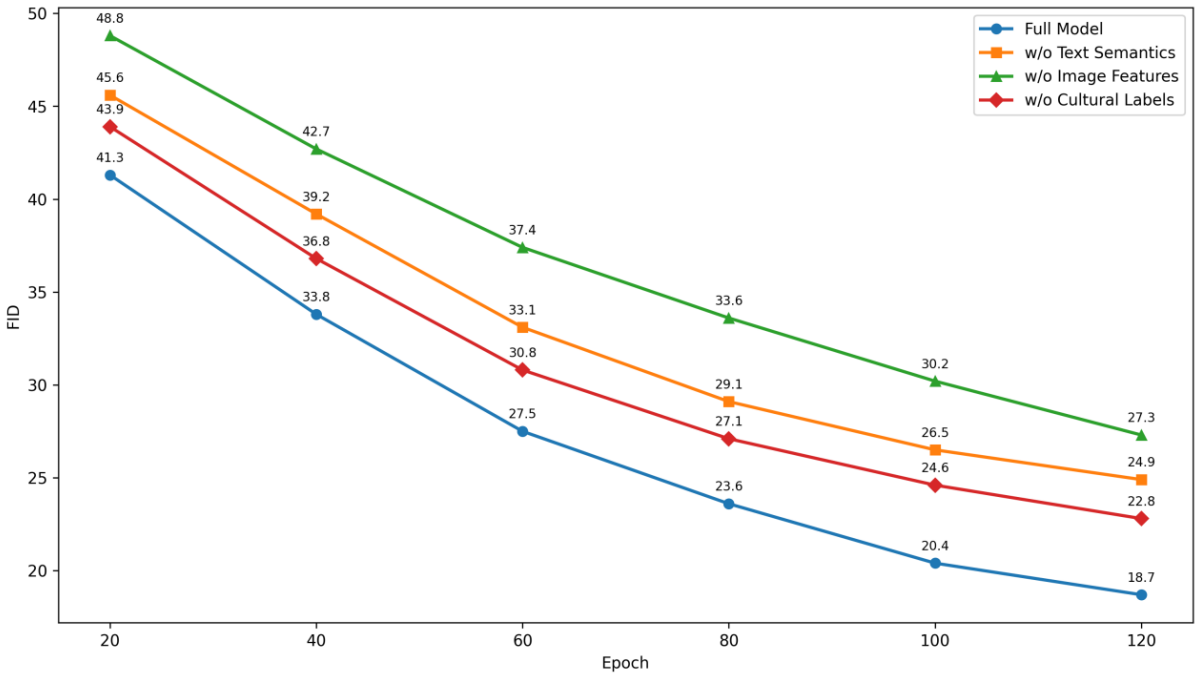

Figure 9: Training FID Trends under Different Ablation Settings

From the perspective of the training process, the multimodal fusion not only improves the final results, but also improves the convergence characteristics of the model. As shown in Figure 9, the FID of the full model continues to decrease from 41.3 to 18.7, always maintaining the optimal performance. After removing the cultural tags, the final FID is 22.8. After removing text semantics and image features, the final FID is 24.9 and 27.3, respectively, and the latter has the smallest decrease, indicating that the lack of image features will significantly weaken the model's ability to fit the visual distribution of complex cultural travel. On the whole, text semantics is responsible for providing theme constraints and cultural narrative orientation, image features determine visual details and composition expression, and cultural labels strengthen regional attributes and style boundaries. The integration of the three can significantly improve the accuracy, stability and design adaptability of digital cultural tourism visual generation.

4.5 Application Case Analysis

In order to further verify the application value of the proposed method in real cultural and tourism communication scenarios, this paper selects four typical tasks, including urban tourism brands, digital posters of scenic areas, cultural and tourism IP images and digital guide interfaces, to carry out case analysis, and comprehensively evaluates them from four dimensions of generation efficiency, cultural element fidelity, user satisfaction and communication suitability, as shown in Table 3.

Table 3: Results of the application case analysis

Application scenarios	Main content generation	Average generation time /s	Fidelity of cultural elements /%	User satisfaction /%	Propagation fitness /%
Urban tourism brand	City main vision, brand extension graphics, publicity cover	6.8	89.3	87.6	88.4
Digital poster of scenic spot	Scenic posters, posters of festival activities, marketing pictures	6.1	91.1	88.7	90.7
Cultural Tourism IP image	Mascots, character drawings, expressions and derivative images	5.9	86.4	85.8	87.1
Digital navigation interface	Navigation home page, column interface, interactive background main vision	6.5	88.6	85.7	89.6
Average value	—	6.33	88.85	86.95	88.95

The results show that the proposed method has good stability and practicability in different types of visual design tasks. The average generation time is 6.8 s, the fidelity of cultural elements reaches 89.3%, the user satisfaction is 87.6%, and the communication adaptability is 88.4%, which indicates that the model can better integrate urban landmarks, regional colors and brand narratives. The digital poster scene of scenic spots has the most obvious advantages in visual impact and scene atmosphere expression, with the fidelity of cultural elements reaching 91.1% and the communication adaptability reaching 90.7%, which is suitable for scenic spot publicity, festival promotion and online marketing. The generation of cultural tourism IP image can complete the output of multi-version scheme in a relatively short time, with an average generation time of 5.9 s. However, due to the high requirements of personalization expression and local symbol details, the user satisfaction rate is 85.8%, slightly lower than that of poster and brand tasks. The digital navigation interface task performs stable in terms of layout coordination and information carrying, and the communication fitness reaches 89.6%, indicating that the method can take into account the requirements of visual generation and interface application. The average generation time of the model is 6.33 s, the average fidelity of cultural elements is 88.85%, and the average user satisfaction is 86.95%. It shows that the method proposed in this paper can not only improve the visual design efficiency of digital cultural tourism, but also achieve a better balance between the accuracy of cultural expression and the adaptability of communication application, which has strong practical promotion value.

5 Conclusion and Prospect

Focusing on the problems existing in the visual image design of digital cultural tourism, such as insufficient efficiency, unstable cultural expression and limited generation control ability, this paper constructs a visual image design method of digital cultural tourism based on artificial intelligence generation model, and forms a relatively complete technical path from three aspects of multi-modal semantic fusion, visual generation mechanism and result feedback optimization. The results show that the proposed method has good feasibility and effectiveness in the visual image generation task of digital cultural tourism. The experimental results show that the average FID of the model on the test set is 18.7, the SSIM is 0.842, the CLIP semantic similarity is 0.781, the fidelity of cultural elements is 88.4%, and the style matching degree is 86.9%. It shows that the method can better realize the mapping from regional culture semantics to visual design results. In the model comparison experiment, the diffusion model outperforms the GAN, VAE and style transfer baselines in terms of generation accuracy, detail performance and design adaptation, with the generation accuracy reaching 89.1% and detail performance reaching 90.3%. In the application case analysis, the average generation time of four typical tasks is 6.33 s, and the average user satisfaction is 86.95%, which shows that the method has strong practical application value in scenes such as urban tourism brand, digital posters of scenic spots, cultural tourism IP image and digital guide interface.

Theoretically, the research of digital cultural tourism visual design is systematically integrated with AIGC generation mechanism, multi-modal semantic representation and computer vision technology, which expands the technical and method boundaries of digital cultural tourism visual image research, and also provides new research ideas for cultural semantic computing, visual generation control and design effect evaluation. From the practical level, the method proposed in this paper can improve the automation level and iterative efficiency of cultural tourism visual design, help alleviate the problems of long cycle, high cost and lack of style continuity in the traditional design process, and provide an operational path for digital cultural tourism brand communication and intelligent visual production.

However, this paper still has some limitations. First, the scale of the current dataset is still limited. Although it contains 12,000 images and 8,500 texts, it still does not adequately cover complex regional cultural scenes. Second, the generalization ability of the model still needs to be improved, and its stability in cross-region, cross-style and cross-media tasks needs to be further verified. Third, the deep semantic understanding of culture still mainly relies on explicit labels and text descriptions, and the modeling of implicit cultural symbols, emotional structures and aesthetic contexts is still insufficient. Fourth, the current generation process is mainly oriented to off-line design tasks, and the ability of real-time interactive generation and dynamic collaborative design still needs to be strengthened. Subsequent research can be carried out from the directions of expanding high-quality multimodal data sets for cultural and tourism, introducing knowledge graphs and semantic reasoning of large models, strengthening controllable generation mechanisms, and building lightweight generation systems for real-time interactive scenarios.

Funding

This research was funded by 2024 Zhejiang Province Education Science Planning Project, China (Project No.2024SCG160); 2024 Zhejiang Province Philosophy and Social Science Planning "Provincial and Municipal Cooperation" Project, China (Project No.

24SSHZ144YB); 2025-2026 Zhejiang Provincial Department of Culture, Radio, Television and Tourism Research and Creative Project: “AI+” Era High-Quality Development Pathways for the Chinese-Style Cultural and Tourism Industry and Zhejiang’s Practices(Project Number: 2025KYZ007).

References

[1] Zhu J, Zhan L, Tan J, et al. Tourism destination stereotypes and generative artificial intelligence (GenAI) generated images[J]. Current Issues in Tourism, 2025, 28(17): 2721-2725.

[2] Chung C, Shin S, Chung N. Satiation of generative AI images[J]. Annals of Tourism Research, 2025, 115: 104054.

[3] Hou L, Min Y, Pan X, et al. Distinguishing AI-generated versus real tourism photos: Visual differences, human judgment, and deep learning detection[J]. Information Processing & Management, 2025, 62(5): 104218.

[4] Wang J. Artificial intelligence (AI) technology in destination advertising: The impact of video-based destination anthropomorphism on destination image[J]. Journal of Destination Marketing & Management, 2025, 35: 100966.

[5] Wang X, Mou N, Zhu S, et al. How to perceive tourism destination image? A visual content analysis based on inbound tourists’ photos[J]. Journal of Destination Marketing & Management, 2024, 33: 100923.

[6] Tan J, Cheng M, Chen J, et al. Multimodal destination image and user engagement: A sequential research design[J]. Tourism Management, 2025, 111: 105209.

[7] Ferracani A, Bertini M, Pala P, et al. Personalized generative storytelling with AI-visual illustrations for the promotion of knowledge in cultural heritage tourism[C]//Proceedings of the 6th workshop on the analySis, Understanding and proMotion of heritAge Contents. 2024: 28-32.

[8] He Z, Su J, Chen L, et al. 'I Recall the Past': Exploring How People Collaborate with Generative AI to Create Cultural Heritage Narratives[J]. Proceedings of the ACM on Human-Computer Interaction, 2025, 9(2): 1-30.

[9] Fu K, Wu R, Tang Y, et al. " Being Eroded, Piece by Piece": Enhancing Engagement and Storytelling in Cultural Heritage Dissemination by Exhibiting GenAI Co-Creation Artifacts[C]//Proceedings of the 2024 ACM designing interactive systems conference. 2024: 2833-2850.

[10] Xu N, Liu Y, Chen Y, et al. ArtifactShow: Incorporating Generative AI into Narrative Visualization for Interactive Cultural Experience[C]//Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 2025: 1-8.

[11] Cardarelli L. From fragments to digital wholeness: An AI generative approach to reconstructing archaeological vessels[J]. Journal of Cultural Heritage, 2024, 70: 250-258.

[12] Cardarelli L. PyPotteryInk: One-step diffusion model for sketch to publication-ready archaeological drawings[J]. Journal of Cultural Heritage, 2025, 74: 300-310.

[13] Li S, Jiang Y, Jing B, et al. AI-based experts' knowledge visualization of cultural heritage: A case study of Terracotta Warriors[J]. Journal of Cultural Heritage, 2025, 72: 81-90.

[14] Kuang Z, Zhang J, Li Y, et al. Preserving architectural heritage in urban renewal: a stable diffusion model framework for automated historical facade generation[J]. npj Heritage Science, 2025, 13(1): 256.

[15] Xiong T, Wang N. Exploring dual pathways for traditional pattern innovation: shape grammar and diffusion models[J]. npj Heritage Science, 2025, 13(1): 639.

[16] Zhou Y, Liu Y, Shao Y, et al. Fine-tuning diffusion model to generate new kite designs for the revitalization and innovation of intangible cultural heritage[J]. Scientific Reports, 2025, 15(1): 7519.

[17] Hu J, Yu Y, Zhou Q. GuidePaint: lossless image-guided diffusion model for ancient mural image restoration[J]. npj Heritage Science, 2025, 13(1): 118.

[18] Zou J, Du Y, Liu G, et al. Generating Chinese intangible cultural heritage images with structure and color awareness[J]. npj Heritage Science, 2025, 13(1): 579.

[19] Xu J, Yan L, Zhang R, et al. A review of the development and application of generative technology in digital museums[J]. npj Heritage Science, 2025, 13(1): 589.