IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1357-1369
Journal homepage: http://iieta.org/journals/ts

User-Generated Images and Artificial Intelligence Large Model-Driven Generation of Tourism Destination Promotional Content

Ningning Xing[1*], Jianhua Liu[2], Xin Ding[2]

[1] College of Culture and Tourism, Zhangzhou Institute of Technology, Zhangzhou 363000, China
[2] College of Tourism, Huaqiao University, Quanzhou 362021, China

Corresponding Author Email: xingningning149147@163.com

https://doi.org/10.18280/ts.430321

Received: 29 March 2026
Revised: 3 June 2026
Accepted: 12 June 2026
Available online: 30 June 2026

Keywords:
controllable image generation, diffusion models, cross-modal feature encoding, geographic semantic constraints, closed-loop optimization, intelligent content generation for cultural tourism

ABSTRACT

The intelligent generation of promotional imagery for tourism has emerged as a critical technological pathway for digital content production in cultural tourism. Existing diffusion-based generation methods, however, struggle to reconcile individualized aesthetic preferences with the geographic semantic fidelity of destination scenes, resulting in a persistent trade-off between subjective stylistic expression and objective spatial structural constraints. Moreover, prevailing generation paradigms rely on static conditional inputs and post-hoc quality filtering, preventing evaluation feedback from being effectively incorporated into the iterative generation process, thereby limiting the adaptability and controllability of generated content. To address these issues, an end-to-end closed-loop image generation framework was constructed, driven jointly by preference and space. In this framework, user aesthetic features and destination geographic topological semantic features were extracted in parallel and adaptively fused through a symmetric cross-gating mechanism for nonlinear heterogeneous feature integration. Efficient guidance of the diffusion model's latent-space evolution was achieved via hypernetwork-based parameter bias injection and spatially biased attention mechanisms, while geographic structural consistency loss was imposed to enforce spatial logical plausibility in the generated outputs. A multi-dimensional joint reward function was further formulated by integrating image realism, geographic fidelity, and preference matching, and a policy gradient algorithm was employed to back-optimize the conditional encoding module, establishing a tightly coupled learning mechanism that cycles through encoding, generation, evaluation, and iterative self-refinement. Experimental results on a self-constructed multi-scenario tourism image dataset demonstrated that the proposed method effectively balanced personalized style rendering with geographic scene constraints, outperforming state-of-the-art algorithms in core generation quality metrics and parameter deployment efficiency, while exhibiting strong cross-scenario adaptability. This work transcends the inherent limitations of static mapping in conventional controllable image generation and establishes a dynamic feedback-driven collaborative generation paradigm for heterogeneous conditioning, offering novel technical directions and theoretical foundations for intelligent content generation in cultural tourism scenarios and for controllable diffusion model optimization under complex conditions.

1. INTRODUCTION

With the rapid advancement of artificial intelligence generation technologies, controllable image generation based on diffusion models has emerged as a central research focus within computer vision and multimedia processing [1, 2]. Research emphasis has progressively shifted from free generation in general scenarios toward domain-specific, precisely controllable generation. The coordination of multi-source heterogeneous prior information and the establishment of intrinsic quality optimization mechanisms during the generation process remain critical scientific challenges that constrain further breakthroughs in controllable generation performance [3, 4]. Against the backdrop of rapid digital transformation in the cultural tourism industry, user-generated visual content has become a core medium for destination branding and scene promotion [5, 6]. Traditional tourism visual content has relied heavily on manual photography and professional post-production, which entails not only high production costs and lengthy workflows but also pronounced stylistic homogenization, rendering such content inadequate for the diverse and personalized content dissemination demands of the new media era [7, 8]. Artificial intelligence generation technologies offer novel solutions for large-scale and customized production of cultural tourism imagery. The construction of an intelligent generation framework that accommodates both subjective user aesthetic preferences and objective geographic semantic features of destinations is expected to advance the theoretical framework of controllable diffusion generation with heterogeneous conditions, while also providing deployable technical support for intelligent cultural tourism content production, thereby bearing

significant academic research value and industrial application potential [9, 10].

Significant technical deficiencies persist in current intelligent image generation systems for cultural tourism scenarios, rendering them inadequate for the production of high-quality tourism promotional content. At the scene adaptation level, existing generation methods predominantly rely on plain text prompts as the sole conditioning input, and the inherent ambiguity of textual semantics prevents models from adequately balancing artistic creation with geographic authenticity, giving rise to an inherent deficiency in which the two performance dimensions constrain each other [11, 12]. Excessive stylistic rendering tends to distort core geographic features such as landmark structures and spatial layouts of destinations, resulting in semantic distortion of scene content. Conversely, generation modes that strictly adhere to realistic views forfeit the capacity for personalized creation and fail to accommodate the aesthetic preferences of different users, thereby precluding the dynamic coordination of subjective creative requirements with objective scene constraints [13, 14]. At the technical architecture level, existing controllable diffusion models predominantly employ unimodal encoding and static feature concatenation for conditional fusion, lacking dynamic coupling modeling capabilities for heterogeneous information such as user aesthetic features and geographic spatial features [15, 16]. Mainstream spatial constraint methods rely on independent external control branches for structural correction, which not only substantially increases model parameter counts and inference overhead but also fails to establish precise binding between pixel-space positions and fine-grained geographic semantics, thereby rendering fine-grained structural control over tourism scenes difficult to achieve [17, 18]. At the optimization mechanism level, existing generation frameworks generally adopt a pipelined workflow characterized by generation followed by filtering, where quality assessment serves merely as a post-hoc evaluation tool. Evaluation signals are unable to be fed back inversely into the feature encoding and latent-space denoising processes, and an end-to-end closed-loop optimization pipeline remains absent. Consequently, models are unable to autonomously adapt to scene characteristics and user preferences, and the overall quality and stability of generated content cannot be iteratively improved [19-24].

To address the aforementioned technical limitations, an end-to-end closed-loop generation framework jointly driven by preference and space is proposed, with systematic innovations established across multiple dimensions. A tightly coupled system integrating encoding, generation, evaluation, and optimization is constructed, enabling a paradigm shift in controllable image generation from static input mapping toward dynamic feedback-driven optimization. Through the design of a symmetric cross-gating fusion structure, adaptive nonlinear fusion of aesthetic features and geographic features is accomplished. Combined with hypernetwork-based parameter bias injection, precise control over heterogeneous conditions is achieved with extremely low parameter overhead. Furthermore, a spatially aware latent diffusion perturbation mechanism is established, which balances global stylistic coherence with fine-grained local geographic structure constraints, thereby ensuring the spatial logical plausibility of generated imagery. By leveraging a multi-dimensional joint reward function and a policy gradient algorithm, autonomous iterative optimization of the conditional encoding module is ultimately realized, leading to sustained improvements in the comprehensive quality of tourism image generation.

The remainder of this study is organized as follows. In Section 2, the overall architecture, core module design, and algorithmic principles of the proposed closed-loop generation framework are elaborated. In Section 3, the effectiveness, parameter efficiency, and cross-scenario adaptability of the proposed method are comprehensively validated through multiple sets of quantitative metric comparisons, modular ablation studies, and visual analyses. In Section 4, the generalization performance and existing limitations of the proposed method are objectively analyzed, and promising directions for future research are discussed. In Section 5, the main research contributions are synthesized, and the core innovations and application values are summarized.

2. CLOSED-LOOP GENERATION FRAMEWORK JOINTLY DRIVEN BY PREFERENCE AND SPACE

2.1 Overall framework architecture

A closed-loop generation framework jointly driven by preference and space is constructed, in which high-precision controllable generation and autonomous optimization of tourism promotional imagery are achieved through three functionally coupled submodules. The overall architecture consists of a joint conditional encoding layer, a geographically semantically guided diffusion decoding layer, and a multi-dimensional preference-aligned reward layer, with all modules forming a complete bidirectional information flow pathway. Heterogeneous user aesthetic data and destination geospatial data are received by the joint conditional encoding layer, where adaptive fusion of heterogeneous information is performed through cross-modal feature learning, and dynamic conditional vectors that integrate both individualized aesthetic characteristics and objective geographic semantic constraints are output. The geographically semantically guided diffusion decoding layer takes the fused conditional vector as the core regulatory basis, and refined controllable denoising operations are executed within the latent space of the diffusion model, thereby achieving high-quality generation of tourism scene imagery. Generation quality is quantitatively evaluated by the multi-dimensional preference-aligned reward layer across multiple dimensions—image realism, geographic fidelity, and user preference matching—and differentiable reward supervision signals are generated to provide guidance for iterative optimization of network parameters. The overall architecture of the closed-loop generation framework jointly driven by preference and space is illustrated in Figure 1.

Two operational modes—training and inference—are distinguished within this framework, thereby enabling a balanced trade-off between optimization performance and inference efficiency. During the training phase, an end-to-end closed-loop learning mechanism encompassing encoding, generation, evaluation, and updating is formed through the three-layer modules, where quality evaluation signals are directly back-propagated to update the parameters of the joint conditional encoding layer. In this manner, generation evaluation is internally embedded within the iterative model optimization process, in contrast to the pipelined post-hoc filtering paradigm characteristic of conventional methods. During the inference phase, only the converged encoding layer and diffusion decoding layer are invoked, while the reward

evaluation branch is discarded, and image generation is accomplished through single-step forward computation, thereby reducing inference overhead while maintaining generation quality. Through this tightly coupled architecture, the inherent limitations of static mapping in controllable generation are overcome, and autonomous evolution of the model's encoding strategy is realized via a dynamic feedback mechanism, effectively enhancing the controllability and adaptive capacity of content generation in complex cultural tourism scenarios.

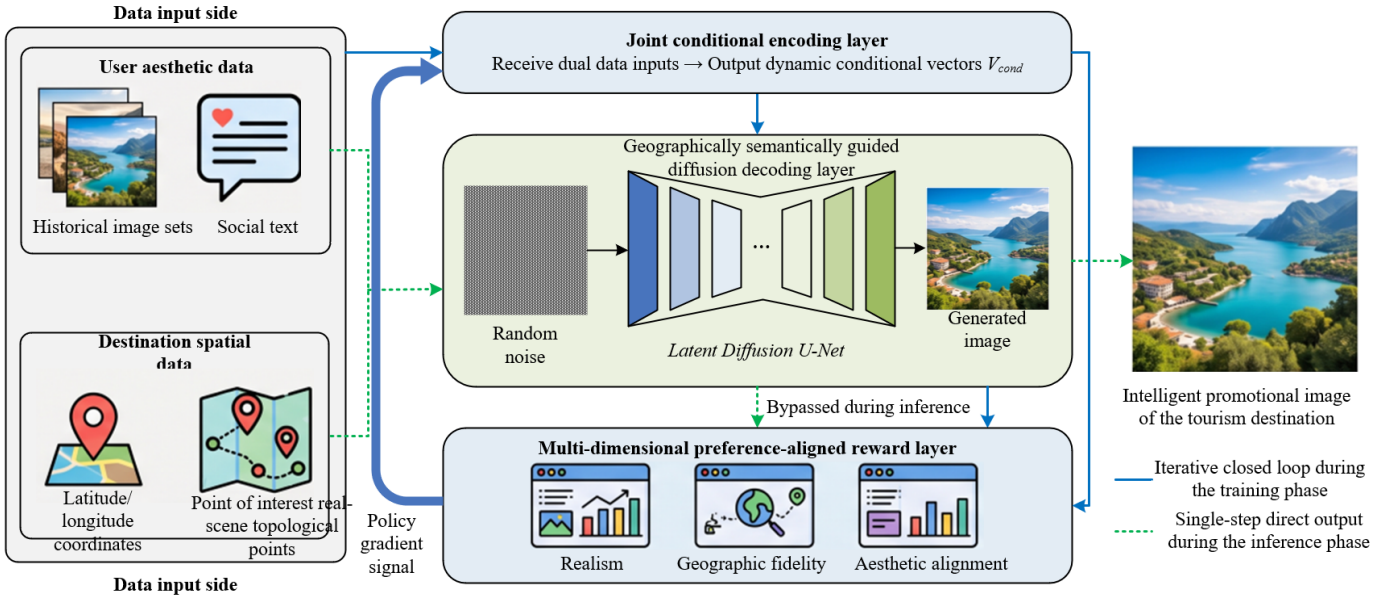

Figure 1. Overall architecture of the closed-loop generation framework jointly driven by preference and space

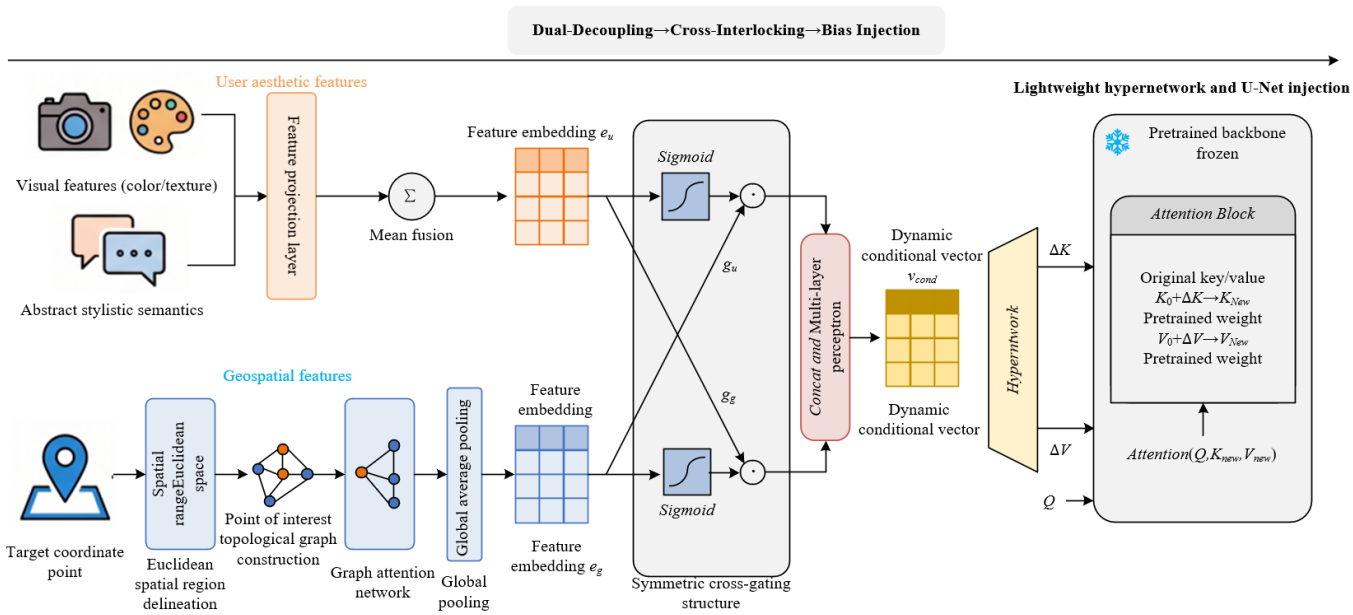

Figure 2. Architecture of the cross-modal symmetric cross-gating joint conditional encoding network

2.2 Cross-modal symmetric cross-gating joint conditional encoding

The cross-modal symmetric cross-gating joint conditional encoding module proposed in this section is designed to accomplish unified representation and adaptive fusion of two types of heterogeneous information—user aesthetic preferences and destination geographic semantics—thereby providing dynamic and precise two-dimensional conditional constraints for the diffusion generation process. The architecture of the cross-modal symmetric cross-gating joint conditional encoding network is illustrated in Figure 2. A dual-branch parallel structure is adopted in the module to enable independent extraction of features from the two sources, through which standardized style representation and spatial semantic representation systems are respectively constructed. In the user preference feature branch, both low-level visual attributes and high-level aesthetic semantics are simultaneously mined from historical user-generated images and social text data, encompassing quantifiable visual features such as color distributions, texture roughness, and proportions of salient regions, as well as abstract stylistic features. The two types of features are dimensionally aligned through dedicated projection layers and uniformly mapped into a d-dimensional latent space, where a global user preference embedding is obtained via element-wise average fusion, with the

representation form given by $E_{style} \in R^d$. In the geographic semantic feature branch, a scene topological graph structure is constructed based on the target geographic location information. An effective spatial range is demarcated with the target coordinates as the center, and point of interest landmarks and functional scenes within this region are treated as graph nodes, with node association weights defined jointly by spatial Euclidean distances and semantic category similarities. Structured spatial features are learned through a lightweight graph attention network, and a fixed-dimensional geographic semantic embedding $E_{geo} \in R^d$ is obtained via global average pooling compression, thereby preserving the spatial layout and scene semantic priors of the tourism destination in their entirety.

To address the limitations of fixed-weight fusion and unidirectional feature modulation, which fail to adapt to dynamic scene constraints, a symmetric cross-gating fusion mechanism is designed to achieve bidirectional adaptive coupling of the two types of heterogeneous features. Through this mechanism, feature weights are mutually modulated by dual gating branches, and the optimal fusion proportion of the two information sources is learned in a data-driven manner. The specific computational expression is given by:

$$v_{cond}=\sigma(W_s E_{style}+b_s)\odot E_{geo}+\sigma(W_g E_{geo}+b_g)\odot E_{style} \quad (1)$$

where, σ denotes the Sigmoid activation function, which generates continuously differentiable soft gating weights; W_s and W_g are learnable weight matrices; b_s and b_g are bias parameters for the respective branches; and $\odot$ represents the element-wise product operation. Unlike the single-constraint mapping logic of conventional unidirectional modulation methods, this mechanism enables mutual constraint and mutual enhancement between the two feature types, with gating weights being dynamically and adaptively adjusted according to the characteristics of the input scene. For strongly structured tourism scenes such as architectural landmarks, the gating response intensity of the geographic semantic branch is elevated, and spatial structure and landmark layout in the generated imagery are preferentially constrained. For weakly structured scenes such as natural landscapes, the user aesthetic branch assumes dominant weighting, and personalized artistic styles are fully rendered while basic scene semantic plausibility is maintained, thereby effectively accommodating the complex and variable generation constraint requirements of cultural tourism scenarios.

To achieve parameter-efficient injection of conditional information, a lightweight hypernetwork is constructed based on the fused dynamic conditional vector v_{cond}, through which fine-grained conditional modulation of the attention mechanism in the pre-trained diffusion model is performed. Incremental bias parameters for the attention layers are dynamically generated by the hypernetwork according to the input conditional features, and minor modifications are applied to the native key and value parameters. The final attention parameter update is formulated as:

$$K'=K+\Delta K,\ V'=V+\Delta V \quad (2)$$

where, K and V denote the intrinsic key and value matrices of the pre-trained U-Net, ΔK and ΔV denote the dynamic conditional biases predicted by the hypernetwork, and K' and V' denote the final attention parameters after the introduction of semantic constraints. All pre-trained parameters of the diffusion backbone network are frozen throughout the entire training process, and external conditional information is introduced solely through a small number of incremental bias parameters, thereby preserving, to the greatest extent possible, the image generation priors learned by the model from general-purpose visual datasets. Furthermore, conditional modulation is applied only to the intermediate-resolution feature layers of the U-Net, while the low-resolution layers are left to maintain global semantic perception capabilities and the high-resolution layers are left to preserve image detail reconstruction capabilities. In this manner, an optimal balance is achieved among controllable generation precision, personalized expressiveness, and parameter computational overhead.

2.3 Spatially aware controllable perturbation mechanism for latent diffusion

To enable refined grounding and regulation of the fused conditional vector within the diffusion latent space, a spatially aware controllable perturbation mechanism for latent diffusion is constructed, through which the model denoising process is collaboratively guided from three dimensions: global style rendering, local semantic binding, and structural constraints. The multi-scale controllable perturbation mechanism for spatially aware latent diffusion is illustrated in Figure 3. A hierarchical perturbation strategy is adopted in this mechanism to overcome the deficiency that single conditional encoding cannot simultaneously maintain global aesthetic consistency and local geographic structural authenticity, thereby enabling synchronous controllability of stylistic expression and spatial layout in tourism imagery. For the unified regulation of global visual style, a channel-wise adaptive style normalization method is employed to perform semantically driven affine transformations on the U-Net feature maps of the diffusion model, where feature distribution patterns are dynamically adjusted to accommodate user aesthetic preferences. The specific computation is formulated as:

$$\widetilde{F}_l=\gamma_l \cdot \frac{F_l-\mu(F_l)}{\sigma(F_l)}+\beta_l \quad (3)$$

where, F_l and $\widetilde{F}_l$ denote the feature maps of the l-th U-Net layer before and after transformation, respectively, and $\mu(F_l)$ and $\sigma(F_l)$ denote the channel-wise mean and standard deviation of the feature maps. The scale parameter γ_l and shift parameter β_l are obtained by mapping the fused conditional vector v_{cond} through a single-layer multilayer perceptron Φ_l, and adaptive normalization parameters are dynamically generated according to user aesthetic features. This approach relies entirely on semantic conditions to drive feature reshaping, without requiring the introduction of external style reference images, thereby fundamentally distinguishing itself from the image transfer logic of conventional adaptive instance normalization. Through this mechanism, semantically controllable unified regulation of global tone, contrast, and texture style is essentially achieved.

Global normalization operations are only capable of homogenized regulation of overall visual style and are unable to impose constraints on fine-grained geographic structures such as local landmarks, building contours, and scene boundaries in tourism scenes. To address this, a learnable spatial bias matrix is embedded within the cross-attention computation process of the diffusion model, through which a

precise associative mapping between latent spatial positions and geographic semantic tokens is established. The optimized attention computation is formulated as:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^T}{\sqrt{d}}+B\right)V \tag{4}$$

where, Q, K, and V denote the query, key, and value matrices of the attention mechanism, respectively; d denotes the feature dimension; and B denotes a spatial bias matrix of dimension $HW\times N$, with HW representing the total number of latent spatial positions and N corresponding to the total number of geographic semantic tokens. The spatial bias matrix is generated by concatenating the fused conditional vector with two-dimensional sinusoidal positional encoding, followed by projection through a two-layer fully connected network, and independent modulation weights are assigned to each spatial position and each geographic semantic token category. Leveraging the temporal characteristics of diffusion denoising, a step-wise scheduling strategy is introduced, wherein spatial bias constraints are enabled only during medium-to-high noise iteration steps to govern the construction of the overall geographic structure of the scene, while bias constraints are disabled during low-noise iteration steps to preserve the model's capacity for reconstructing high-frequency image details. Through this approach, an effective balance between geographic structural fidelity and visual detail richness is achieved.

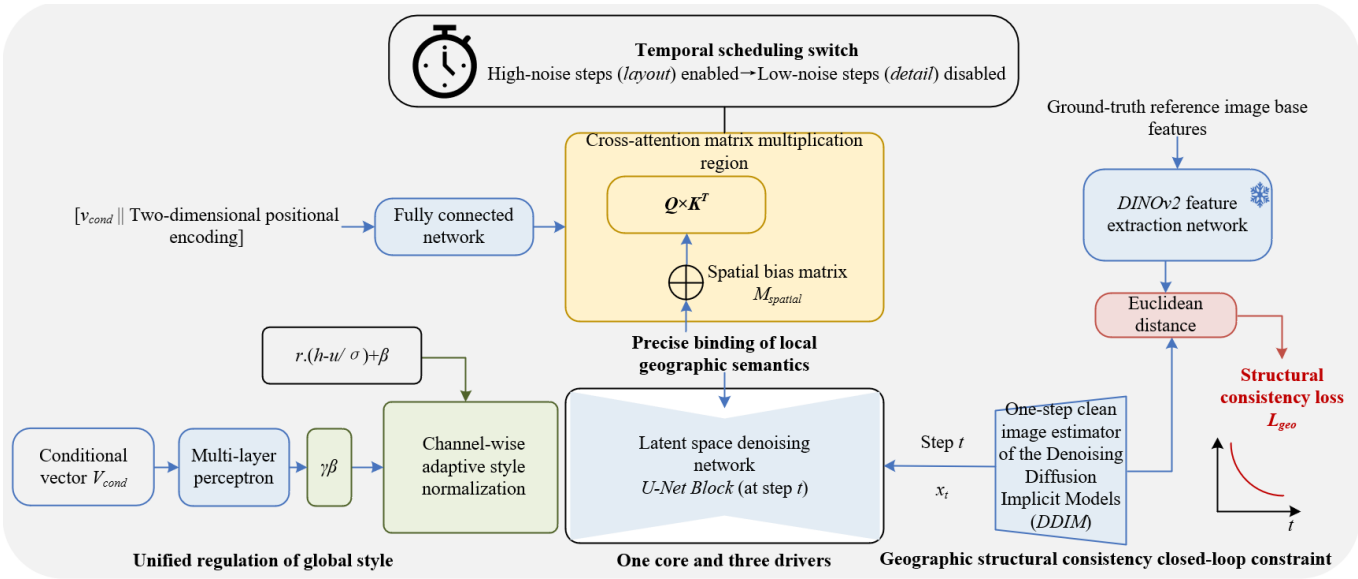

Figure 3. Multi-scale controllable perturbation mechanism for spatially aware latent diffusion

To further solidify the geographic authenticity of scene spatial layouts and compensate for the limitations of soft attention-based constraints, a geographic structural consistency loss function based on self-supervised features is constructed, through which strong supervisory constraints are imposed at each iteration step of the diffusion process. A clean image $\hat{x}_0$ at the current iteration stage is estimated in real time from the noisy latent variables via the inversion mechanism of the Denoising Diffusion Implicit Models (DDIM), and layout constraints are enforced by aligning the deep structural features of the generated image with those of the ground-truth scene. The loss function is defined as:

$$L_{geo}=\|DINOv2(\hat{x}_0)-DINOv2(x_{ref})\|_2 \tag{5}$$

where, $DINOv2(\,)$ denotes the self-supervised feature extraction operation, and x_{ref} denotes the mean feature baseline of ground-truth images of the target tourism destination, which is used to represent the intrinsic geographic structural prior of the scene. The pre-trained *DINOv2* features exhibit strong spatial structure perception capabilities and robustness to variations in texture and color, enabling accurate capture of core geographic information such as scene layout and relative landmark positions. To balance structural constraints with generation diversity, a dynamic loss weight that decays linearly with noise intensity is introduced, with the weight coefficient given by $\lambda_{geo}=1-t/T$, where t denotes the current iteration step and T denotes the total number of denoising steps. A larger loss weight is assigned during high-noise stages to prioritize the correction of overall scene layout, while constraint intensity is weakened during low-noise stages to ensure sufficient creative freedom for the model.

The three-tier perturbation mechanism forms a progressively layered, soft–hard integrated latent space regulation system, through which refined control over the entire diffusion generation process is achieved. Channel-wise normalization accomplishes unified shaping of global aesthetic style, spatial bias attention enables precise binding of local geographic semantics, and structural consistency loss imposes closed-loop constraints on scene spatial logic at the feature level. Through the synergistic action of the three components, the latent-space evolution of the diffusion model is guided simultaneously by both user aesthetic preferences and destination geographic priors, fundamentally addressing the imbalance between stylistic rendering and real-scene structural fidelity in cultural tourism image generation.

2.4 Multi-reward guided closed-loop optimization of the conditional encoder

To achieve dynamic and iterative improvement in generation quality, a reinforcement learning mechanism is introduced to establish an end-to-end closed-loop optimization system, through which the conditional encoding strategy is

adaptively updated under the reverse constraint of generation performance. Conventional optimization approaches for diffusion-based generation typically operate directly on the denoising network, where the step-wise latent-space evolution process tends to induce gradient oscillation and training redundancy. In this work, the complete tourism image generation pipeline is modeled as an episodic Markov decision process, through which the complex temporal optimization logic is effectively simplified. Within a single decision episode, the input user aesthetic features and geographic semantic features constitute the system state, the fused dynamic conditional vector serves as the agent's decision action, and the comprehensive quality score of the final generated image serves as the episodic supervisory reward. Through this modeling approach, the continuous optimization problem of multi-step denoising is transformed into a single-step conditional encoding decision problem, thereby significantly compressing the gradient computation dimension, suppressing gradient variance during training, and enhancing overall iterative stability. The multi-reward guided closed-loop optimization flow of the conditional encoder is illustrated in Figure 4.

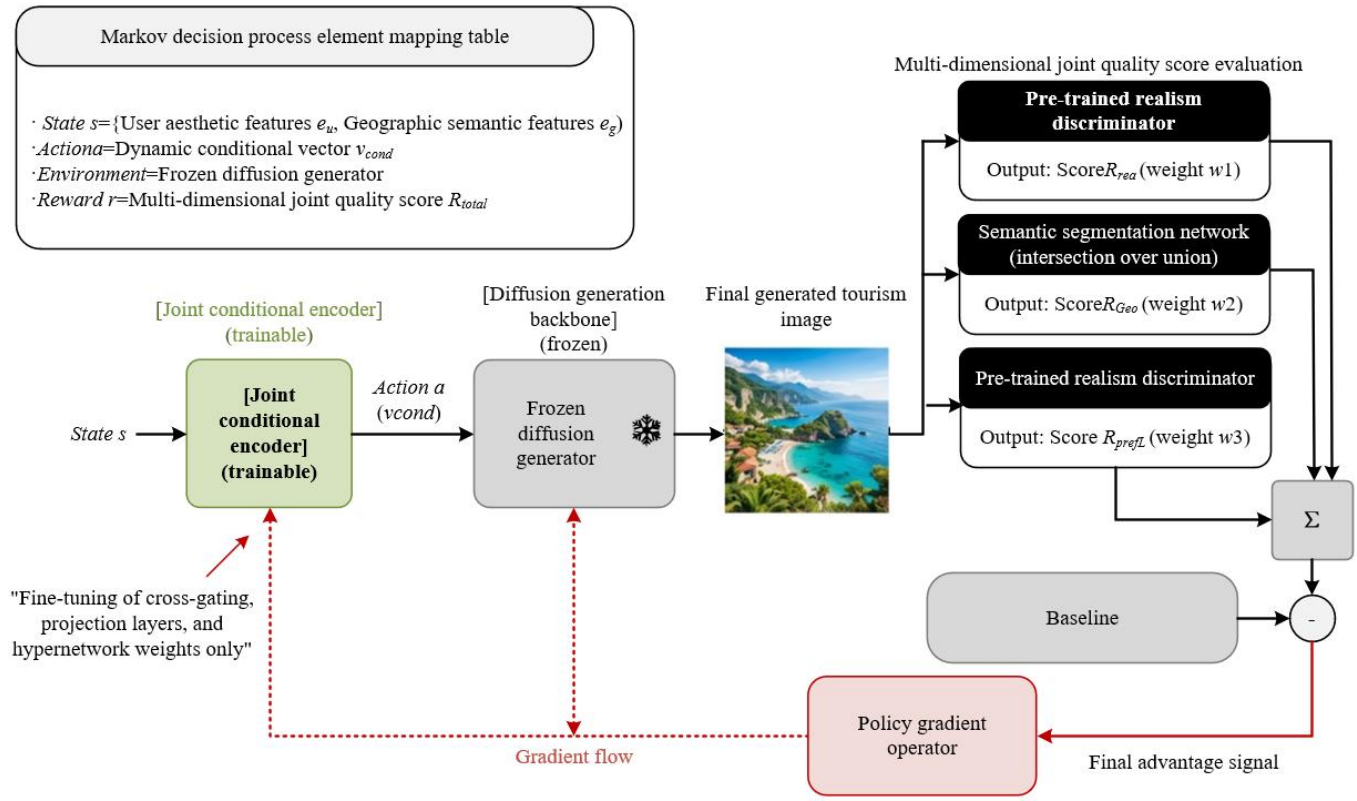

Figure 4. Multi-reward guided closed-loop optimization flow of the conditional encoder

To comprehensively quantify the overall generation quality of cultural tourism imagery, a multi-dimensional fused reward function system is constructed, through which quantitative supervisory signals are established across three dimensions: image realism, geographic semantic matching, and user aesthetic alignment. Each reward component is first normalized to a unified numerical interval via min-max scaling, thereby eliminating weight bias caused by dimensional discrepancies. The realism reward is computed by evaluating the distributional discrepancy between generated images and real cultural tourism images using a pre-trained discriminator, with the formulation given by:

$$R_{real}=\log\left(D(\hat{x}_0)\right) \tag{6}$$

The geographic fidelity reward quantifies the degree of scene structural matching via semantic segmentation intersection over union, expressed as:

$$R_{geo}=mIoU(Seg(\hat{x}_0),Seg_{target}) \tag{7}$$

The user preference alignment reward measures the correspondence between generated content and user aesthetic style through feature cosine similarity, expressed as:

$$R_{pref}=\cos\left(Encoder_{img}(\hat{x}_0),\mathrm{E}_{style}\right) \tag{8}$$

The three reward components are combined through weighted fusion to obtain the final joint reward signal, with the overall formulation given by:

$$R=\lambda_1 R_{real}+\lambda_2 R_{geo}+\lambda_3 R_{pref} \tag{9}$$

where, λ_1, λ_2, and λ_3 denote the balancing weights for each reward component, respectively. A user preference prediction model is pre-trained in an offline manner, and its parameters remain fixed throughout the closed-loop training process, enabling the generation of stable and reliable aesthetic matching supervisory signals while avoiding reward fluctuation issues caused by dynamic training.

Based on the constructed joint reward signal, a policy gradient algorithm with baseline subtraction is employed to iteratively update the parameters of the conditional encoding network, with the optimization objective defined as the maximization of the expected reward. The policy gradient is computed as:

$$\nabla_\theta J(\theta)=E_{\hat{x}_0\sim p_\theta(\cdot|v_{cond})}\left[(R(\hat{x}_0)-b)\cdot\nabla_\theta \log p_\theta(\hat{x}_0|v_{cond})\right] \quad (10)$$

where, θ denotes all trainable parameters of the joint conditional encoding layer, and b denotes the baseline parameter computed as the mean reward over historical iterations. Through baseline deviation correction, gradient estimation bias is further reduced and training convergence efficiency is enhanced. The reparameterization trick is introduced to facilitate gradient computation in the continuous action space, where the stochastic sampling process of the conditional vector is transformed into a deterministic mapping superimposed with controllable noise, thereby effectively resolving the difficulty of gradient estimation in continuous decision spaces.

In terms of the parameter update paradigm, a localized fine-tuning optimization strategy is adopted, wherein only the parameters of the symmetric cross-gating fusion module, feature projection layers, and hypernetwork are updated via back-propagation, while the parameters of the diffusion backbone network, feature extraction networks, and reward evaluation models remain frozen throughout training. Through this optimization strategy, the general image generation priors of the pre-trained diffusion model are fully preserved, and only the fusion and injection strategies for heterogeneous conditions are targeted for optimization. In this manner, sustained iterative improvement in generation quality is achieved while parameter efficiency and training stability are simultaneously maintained, ultimately establishing an endogenous closed-loop optimization paradigm encompassing encoding, generation, evaluation, and updating.

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 Experimental setup

Based on the specific characteristics of image generation tasks in cultural tourism scenarios, a self-constructed multi-scenario tourism image dataset was established for model training and performance evaluation. The dataset encompassed three major tourism scene categories—landmark buildings, natural landscapes, and urban streetscapes—comprising data samples from a total of 86 popular tourism destinations both domestically and internationally. The data composition included destination ground-truth reference images, user-generated content tourism images, and corresponding geographic semantic annotations, comprehensively covering both strongly structured landmark scenes and weakly structured natural scenes. A standardized preprocessing pipeline was uniformly applied to the dataset, including image size normalization, data cleaning, and semantic label alignment. The dataset was partitioned into training, validation, and test sets at a ratio of 7:1:2. The training set was used for iterative model parameter updates, the validation set was employed for hyperparameter tuning, and the test set was reserved for final performance evaluation and generalization capability validation, thereby ensuring the objectivity and reliability of the experimental results.

To comprehensively validate the performance advantages of the proposed method, four categories of state-of-the-art technical approaches were selected to establish comparison baselines, covering four major technical trajectories: text-native generation, style fine-tuning, geographic structure constraints, and static feature fusion. The baselines were configured as follows. The baseline text-native generation method employed the native Stable Diffusion model with text prompts as the sole conditioning input, achieving general-purpose image generation without additional constraints. The style fine-tuning method performed lightweight fine-tuning of the pre-trained diffusion model via low-rank adaptation, learning a fixed aesthetic style from user-generated content data to achieve stylized tourism image generation. The geographic constraint method incorporated the ControlNet semantic segmentation mask constraint mechanism, forcing the model to match the geographic scene structure of the destination to ensure spatial authenticity of the generated content. The static fusion method retained the complete network architecture of the proposed approach while removing the multi-reward closed-loop optimization branch, and adopted direct feature concatenation for dual-source feature fusion, serving to validate the effectiveness of the closed-loop optimization mechanism.

Stable Diffusion-v1.5 was adopted as the pre-trained backbone model, and all experiments were implemented using the PyTorch deep learning framework. The hardware environment was configured with a single NVIDIA RTX 4090 graphics processing unit with 24 GB of video memory. The AdamW optimizer was employed during training, with a base learning rate set to 1e-4, a weight decay coefficient set to 5e-5, a batch size set to 8, and a total of 200 training epochs. The hypernetwork and feature projection layer parameters were randomly initialized, while the parameters of the diffusion backbone network, DINOv2 feature extraction network, semantic segmentation network, and preference prediction network remained frozen throughout training to ensure that the pre-trained priors were not disrupted. During the closed-loop optimization phase, the reward weight coefficients were set to $\lambda_1 = 0.3$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.3$, balancing image realism, geographic fidelity, and user preference alignment.

3.2 Overall performance comparison experiments

To comprehensively validate the overall performance advantages of the proposed framework in tourism image generation tasks, the complete model was compared horizontally against the four baseline methods across all evaluation metrics. The experimental results are presented in Table 1.

As shown by the quantitative results, the proposed method significantly outperformed all baseline models across all core evaluation metrics. Compared with the native diffusion model, the Fréchet Inception Distance (FID) score was reduced by 13.61 and the inception score was increased by 8.51, demonstrating that the closed-loop optimization and feature fusion mechanisms substantially enhance the realism and naturalness of the generated imagery. In comparison with the low-rank adaptation style fine-tuning model, the proposed method achieved a 0.329 improvement in the geographic semantic mean intersection over union metric while maintaining high preference matching, effectively addressing the issue that conventional style fine-tuning methods overemphasize aesthetic rendering at the expense of geographic scene structure. Relative to the ControlNet geographic constraint model, the proposed method achieved higher geographic fidelity and user preference alignment with less than one-quarter of the parameter count, circumventing the redundancy of parameter overhead and the limitation of monolithic style expression inherent in conventional external

control branches. Compared with the static fusion model, the closed-loop optimization mechanism further elevated all metrics across the board, confirming that dynamic feedback-based iterative refinement consistently improves the conditional encoding strategy, thereby achieving synergistic enhancement of both geographic authenticity and personalized style. The subjective mean opinion scores were highly consistent with the objective metrics, and the images generated by the proposed method exhibited clear advantages in scene plausibility, style compatibility, and visual aesthetic quality.

Table 1. Overall performance comparison results of different methods

Method	Fréchet Inception Distance (FID)↓	Inception Score↑	Mean Intersection over Union↑	Preference Cosine Similarity↑	Trainable Parameters (M)↓	Mean Opinion Score↑
Native Stable Diffusion	23.76	22.35	0.521	0.613	0	3.21
Low-Rank Adaptation Style Fine-Tuning	18.42	25.68	0.547	0.826	12.36	3.68
ControlNet Geographic Constraint	16.85	26.12	0.783	0.652	38.72	3.75
Static Fusion Ablation Model	14.28	27.45	0.812	0.841	8.65	3.92
Proposed Full Method	10.15	30.86	0.876	0.915	8.71	4.46

Table 2. Core module ablation experimental results

Model Variant	Symmetric Cross-Gating	Spatial-Aware Perturbation	Closed-Loop Optimization	Fréchet Inception Distance (FID)↓	Mean Intersection over Union↑	Preference Cosine Similarity↑	Mean Opinion Score↑
1. Baseline Model	×	×	×	23.76	0.521	0.613	3.21
2. Direct Feature Concatenation	×	×	×	20.14	0.605	0.725	3.45
3. Symmetric Cross-Gating Encoding	√	×	×	16.32	0.743	0.838	3.78
4. Dual-Module Combination	√	√	×	12.87	0.835	0.862	4.05
5. Proposed Full Model	√	√	√	10.15	0.876	0.915	4.46

Table 3. Model parameter efficiency comparison results

Fine-Tuning Method	Trainable Parameters (M)↓	Convergence Epochs↓	Fréchet Inception Distance (FID)↓	Mean Intersection over Union↑
Full Fine-Tuning	892.35	160	15.68	0.791
Standard Low-Rank Adaptation Fine-Tuning	12.36	120	18.42	0.547
Conventional Adapter Fine-Tuning	25.84	100	13.56	0.804
Proposed Method	8.71	80	10.15	0.876

3.3 Core module ablation experiments

To validate the independent effectiveness and synergistic gains of each core innovative module, five progressive model variants were designed for ablation experiments, through which the performance contributions of the three major modules—symmetric cross-gating fusion, spatially aware diffusion perturbation, and closed-loop reward optimization—were quantitatively analyzed. The experimental results are presented in Table 2.

The experimental results demonstrate that each module contributes positive performance gains to the model, and the combination of multiple modules exhibits significant synergistic optimization effects. Compared with the baseline model, simple feature concatenation fusion only marginally improved generation performance, with limited capacity for heterogeneous feature integration. After the introduction of the symmetric cross-gating fusion mechanism, both preference matching and geographic fidelity exhibited substantial improvements, confirming that the bidirectional adaptive gating structure effectively resolves fusion conflicts between heterogeneous features and accommodates the dual-constraint generation requirements of cultural tourism scenarios. Following the addition of the spatially aware latent diffusion perturbation mechanism, the model's capacity for geographic structural constraint was further enhanced, with the mean intersection over union metric improved by 0.092, indicating that hierarchical style regulation and spatial constraint strategies effectively regularize the diffusion denoising process and improve scene layout plausibility. After the final incorporation of the closed-loop reward optimization mechanism, comprehensive performance improvements were achieved across all metrics, confirming that dynamic feedback-based optimization consistently refines the conditional encoding strategy and breaks through the performance bottleneck of static generation paradigms.

Through synergistic interaction, the modules collectively establish a complete high-precision controllable generation system.

3.4 Parameter efficiency comparison experiments

To validate the lightweight advantages of the proposed conditional injection scheme, the hypernetwork-based bias injection method was compared against three mainstream fine-tuning approaches—full fine-tuning, standard low-rank adaptation, and conventional adapter—in terms of parameter efficiency and generation performance. The experimental results are presented in Table 3.

As shown by the experimental results, the proposed method achieves optimal overall performance across parameter count, convergence speed, and generation accuracy. Full fine-tuning requires updating all model parameters, incurring substantial parameter overhead and extended convergence cycles, making it unsuitable for lightweight deployment scenarios. Standard low-rank adaptation fine-tuning offers relatively manageable parameter overhead but is only capable of style adaptation without effective constraint of geographic semantic structure, resulting in poor scene fidelity performance. Conventional adapter fine-tuning provides a certain degree of structural constraint capability but suffers from excessive parameter redundancy and suboptimal convergence efficiency. Through the hypernetwork-based dynamic bias injection strategy, only a small number of network parameters are fine-tuned in the proposed method, with the trainable parameter count significantly lower than that of all comparison methods, while also requiring the fewest convergence epochs. Despite its lightweight nature, the model maintains optimal image realism and geographic fidelity, fully demonstrating that this parameter injection paradigm enables precise conditional control with extremely low computational overhead, offering substantial value for engineering deployment and edge computing scenarios.

3.5 Cross-scenario adaptive capability validation experiments

To validate the scene-adaptive fusion characteristics of the symmetric cross-gating mechanism, comparative tests were conducted across three typical tourism scene categories—landmark buildings, natural landscapes, and urban streetscapes—with model performance and mean gating weights statistically analyzed for each scene. The experimental results are presented in Figure 5 and Table 4.

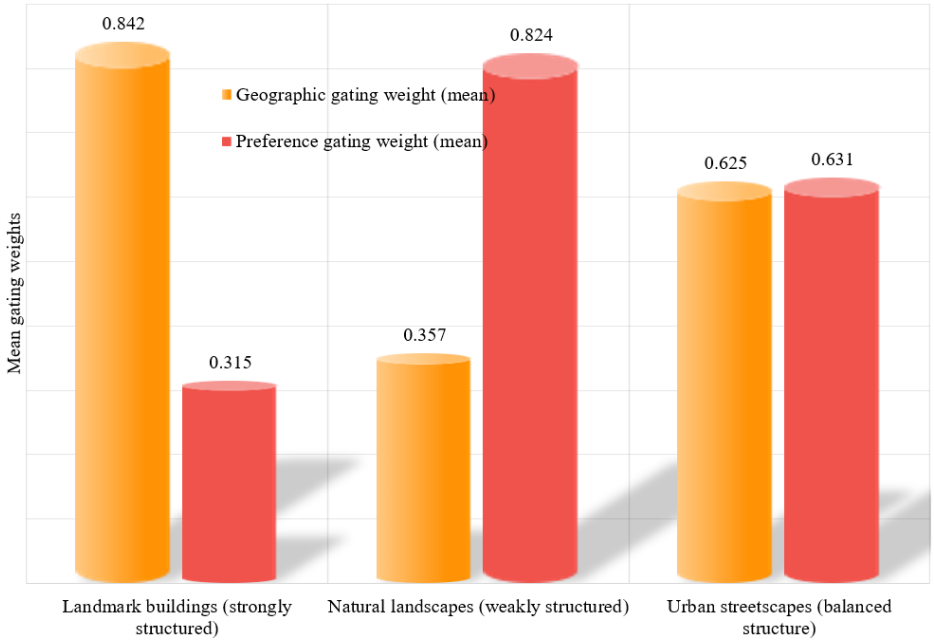

Figure 5. Mean gating weights

Table 4. Cross-scenario adaptive performance test results

Scene Category	Mean Intersection over Union↑	Preference Cosine Similarity↑
Landmark Buildings (Strongly Structured)	0.902	0.886
Natural Landscapes (Weakly Structured)	0.831	0.942
Urban Streetscapes (Balanced Structure)	0.895	0.912

The experimental results fully demonstrate the dynamic adaptive fusion capability of the model. In strongly structured scenes such as landmark buildings, the geographic feature gating weight is automatically elevated by the model, and the accuracy of landmark contours, spatial layouts, and architectural structures is prioritized, thereby achieving extremely high geographic fidelity. In weakly structured scenes such as natural landscapes, the preference gating weight assumes a dominant role, and stringent structural constraints are relaxed by the model, allowing full expression of personalized aesthetic generation capabilities and significantly improving user preference matching. For urban streetscape scenes, where structural and stylistic demands are balanced, the two gating weights tend toward equilibrium, enabling balanced optimization of geographic authenticity and stylistic personalization. These results confirm that the symmetric cross-gating mechanism autonomously adjusts the fusion ratio of dual-source features according to inherent scene

characteristics, thoroughly overcoming the scene adaptation deficiencies of fixed-weight fusion schemes and exhibiting robust adaptability across the full spectrum of cultural tourism scenarios.

3.6 Closed-loop evolution convergence analysis experiments

To investigate the iterative convergence characteristics of the multi-reward closed-loop optimization mechanism, the comprehensive reward values and core generation metrics of the model were statistically analyzed across different training iterations, and the dynamic evolution patterns of the closed-loop optimization were examined. The experimental results are presented in Table 5.

As can be observed from the iterative evolution trends, the comprehensive reward value of the model steadily increased with training epochs, and all generation performance metrics were simultaneously optimized, with convergence approached after 150 iterations and optimal steady-state performance reached at 200 iterations. During the early training stages, the model exhibited substantial improvements in reward and performance, and the closed-loop feedback rapidly corrected inherent deficiencies in the static encoding strategy, leading to swift enhancement of image generation quality. During the later training stages, the optimization rate gradually slowed and the metrics leveled off, indicating that the conditional encoding strategy had completed adaptive iteration and reached an optimal matching state between input conditions and the generation process. Compared with the static model without closed-loop optimization, the converged full model achieved significant improvements across all metrics, demonstrating that the reward-guided closed-loop optimization mechanism is not merely a superficial performance fine-tuning process, but rather achieves autonomous evolution of the encoding strategy through sustained iteration, effectively breaking through the performance ceiling of static generation frameworks.

Table 5. Closed-loop optimization iterative convergence performance results

Iteration Epoch	Comprehensive Reward Value↑	Fréchet Inception Distance (FID)↓	Mean Intersection over Union↑	Preference Cosine Similarity↑
0 (Static Model)	0.725	14.28	0.812	0.841
50	0.793	12.65	0.834	0.868
100	0.856	11.32	0.857	0.892
150	0.894	10.48	0.871	0.908
200	0.901	10.15	0.876	0.915

Figure 6. Implementation effect of intelligent generation of tourism promotional images driven by preference–geography dual-conditions

3.7 Visual qualitative analysis

To validate the collaborative control capability of the proposed framework over user preference expression and destination geographic authenticity under real-world cultural tourism image conditions, visual analyses of the implementation process and result discrepancies across multiple methods were further conducted. As shown in Figure 6, the model is capable of extracting aesthetic cues—such as fresh colors, open compositions, and vivid photography—from user-generated content images, while simultaneously obtaining geographic semantic constraints from destination

elements including plateau lakes, forested mountains, ancient city towers, rammed-earth foundations, and ethnic costumes. Throughout the diffusion generation process, the transformation from high-noise latent variables to clear promotional imagery is progressively accomplished through structural shaping, style rendering, and detail refinement. The final generated results visually preserve lake shorelines, mountain contours, ancient architectural forms, and folk character features, indicating that preference conditions and geographic conditions do not undergo simple superposition but rather form effective coupling during the generation process. Figure 7 further demonstrates that generation quality across different methods exhibits stable gradient disparities in both scene categories. The native Stable Diffusion results possess certain visual appeal but suffer from insufficient destination recognizability. Low-rank adaptation style fine-tuning enhances color and atmosphere at the expense of spatial structural consistency. ControlNet geographic constraints improve contour preservation yet remain inadequate in style adaptation. The static fusion model achieves relatively balanced results, whereas the proposed full method attains the highest geographic fidelity level in both natural landscape and cultural landscape scenes. These qualitative findings are consistent with the quantitative experimental results: the proposed method achieves an FID of 10.15, an inception score of 30.86, a mean intersection over union of 0.876, a preference cosine similarity of 0.915, and a mean opinion score of 4.46, representing a 13.61 reduction in FID, an 8.51 increase in the inception score, a 0.355 improvement in the mean intersection over union, and a 0.302 enhancement in preference similarity over the native Stable Diffusion baseline, while introducing only 8.71 M trainable parameters—significantly lower than the 38.72 M parameter scale of ControlNet. These results collectively demonstrate that the proposed closed-loop generation framework jointly driven by preference and space is capable of simultaneously enhancing image realism, geographic structural fidelity, and user aesthetic alignment with low parameter overhead, thereby providing reliable support for personalized, high-fidelity, and large-scale intelligent generation of tourism destination promotional content.

Figure 7. Comparison of generation result discrepancies and geographic fidelity effects across multiple methods

4. DISCUSSION

The closed-loop generation framework jointly driven by preference and space proposed in this work exhibits excellent model compatibility and cross-scenario generalization capability. A modularly decoupled architecture is adopted in the overall framework, wherein the conditional encoding, spatial perturbation regulation, and closed-loop optimization submodules are independent of the diffusion backbone network, enabling adaptation to various mainstream pre-trained diffusion foundations without requiring large-scale restructuring or fine-tuning of the main model architecture. At the level of cultural tourism application scenarios, the framework is capable of adaptively accommodating diverse regional landscapes and scene structural characteristics. Both precise constraint of spatial proportions and layout logic in highly structured landmark buildings and flexible adaptation to personalized stylistic creation requirements in weakly structured scenes—such as natural landscapes and urban streetscapes—are achieved, demonstrating universal regulatory capacity across diversified tourism scenes. Furthermore, the model dynamically adjusts the fusion weights of dual-source features according to the aesthetic characteristics of different users, accommodating differentiated user-generated content creation styles such as fresh, vintage, and textured aesthetics. From the perspective of technical transferability, the heterogeneous feature adaptive fusion and feedback-driven iterative optimization mechanisms constructed in this work are not limited to cultural tourism image generation scenarios but can be extended to multiple cross-modal generation tasks, including controllable remote sensing image generation and customized scene-oriented

visual creation, thereby offering substantial general research value and broad extension potential.

Although the proposed method achieves synergistic optimization of geographic fidelity and preference matching in conventional tourism scenarios, certain performance limitations persist in extreme and niche scenes. For niche cultural tourism scenes with scarce sample accumulation, the limited scene prior information prevents the model from accurately capturing exclusive spatial structures and semantic features, and the adaptive regulation precision of the gating fusion mechanism is correspondingly diminished, resulting in slight deviations in the geographic detail authenticity of generated content. When users exhibit extreme aesthetic preferences characterized by highly artistic or surreal styles, a trade-off conflict arises between the demand for intensive stylistic rendering and the constraint of geographic structural fidelity, and the model struggles to achieve absolute equilibrium between the two optimization objectives, frequently giving rise to issues such as localized weakening of scene structure and fine-grained semantic deviations. Furthermore, under extreme lighting conditions—including backlighting, overexposure, and nighttime scenes—the distribution of low-level visual features in the imagery undergoes shifts, and the regulatory stability of the latent-space constraint mechanism is reduced, leading to unnatural light transitions and localized detail distortions in the generated imagery. Consequently, there remains room for improvement in the model's robustness to complex and extreme environments.

This study offers novel theoretical insights and technical paradigm support for the field of controllable diffusion generation. Existing controllable generation research has largely relied on static conditional inputs and open-loop generation logic. In contrast, through the construction of a closed-loop encoding-side optimization system in this work, it has been verified that multi-dimensional quality feedback signals can endogenously drive the autonomous iterative evolution of conditional encoding strategies, thereby breaking through the performance bottleneck of static mapping generation paradigms and advancing the theoretical framework of controllable generation with multi-heterogeneous-condition collaborative constraints and multi-objective joint optimization. The hierarchical spatially aware perturbation and bidirectional adaptive feature fusion strategies also provide novel technical pathways for resolving multi-source condition conflicts and achieving fine-grained scene semantic control. Future research can further integrate three-dimensional geographic information data to strengthen precise constraints on the three-dimensional spatial structure of tourism scenes, and incorporate multi-round human–computer interaction mechanisms to enable dynamic iterative updating of user aesthetic preferences. Additionally, the closed-loop collaborative optimization paradigm can be extended to multi-dimensional digital content generation tasks—such as cultural tourism short videos and panoramic imagery—thereby further promoting the refinement and deployment of intelligent content generation technology systems for smart cultural tourism.

5. CONCLUSION AND FUTURE DIRECTIONS

To address the industry and technical challenges in intelligent generation of tourism destination promotional imagery—specifically, the difficulty of coordinately controlling aesthetic preferences and geographic semantics, and the lack of intrinsic optimization mechanisms during the generation process—a closed-loop controllable generation framework jointly driven by preference and space was constructed in this study. Unified modeling and adaptive fusion of multi-source heterogeneous cultural tourism information were accomplished, and dynamic coupling of user aesthetic features and geospatial features was realized through a symmetric cross-gating mechanism. Global style regulation and local geographic structure constraints were achieved via a spatially aware latent perturbation strategy, and a closed-loop iterative optimization mechanism at the conditional encoding side was established through multi-dimensional reward supervisory signals. The results of comparative and ablation experiments demonstrated that the proposed method effectively balanced image realism, geographic fidelity, and user preference matching, and exhibits superior overall generation performance, parameter deployment efficiency, and cross-scenario adaptive capability relative to conventional controllable generation methods. This work breaks through the inherent limitations of traditional static conditional mapping generation paradigms, enriches the theoretical and methodological framework of heterogeneous-condition collaborative controllable generation, and provides a reliable technical solution for large-scale, high-quality intelligent content production in the cultural tourism domain.

Building upon the existing research foundation, future work can be further extended and deepened across multiple dimensions. Three-dimensional geospatial information can be integrated to establish a three-dimensional scene constraint mechanism, thereby resolving spatial structural biases inherent in two-dimensional image generation and enabling immersive cultural tourism content generation. Additionally, real-time human–computer interactive iteration strategies can be introduced to dynamically capture and update user aesthetic preferences, further enhancing the personalized adaptation capability of the model. Furthermore, the closed-loop optimization generation paradigm can be extended to dynamic cultural tourism promotional content generation tasks—such as short videos and panoramic imagery—while optimizing model robustness under extreme lighting conditions and niche scenes, thereby continuously advancing the deep application and innovation of controllable generation technologies in the digital transformation of smart cultural tourism.

ACKNOWLEDGEMENTS

This work is supported by the Zhangzhou Institute of Technology (Grant No.: ZZYB2209).

REFERENCES

[1] Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869. https://doi.org/10.1109/TPAMI.2023.3261988

[2] Cao, H., Tan, C., Gao, Z., et al. (2024). A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering, 36(7): 2814-2830. https://doi.org/10.1109/TKDE.2024.3361474

[3] Jiang, R., Zheng, G.C., Li, T., Yang, T.R., Wang, J.D., Li, X. (2024). A survey of multimodal controllable diffusion models. Journal of Computer Science and Technology, 39(3): 509-541. https://doi.org/10.1007/s11390-024-3814-0
[4] Bie, F., Yang, Y., Zhou, Z., et al. (2025). RenAIssance: A survey into AI text-to-image generation in the era of large model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3): 2212-2231. https://doi.org/10.1109/TPAMI.2024.3522305
[5] Oliveira, T., Araujo, B., Tam, C. (2020). Why do people share their travel experiences on social media? Tourism Management, 78: 104041. https://doi.org/10.1016/j.tourman.2019.104041
[6] Nguyen, T.T.T., Tong, S. (2023). The impact of user-generated content on intention to select a travel destination. Journal of Marketing Analytics, 11(3): 443-457. https://doi.org/10.1057/s41270-022-00174-7
[7] Guerreiro, M.M., Pinto, P.S., Ramos, C., et al. (2024). The online destination image as portrayed by the user-generated content on social media and its impact on tourists' engagement. Tourism & Management Studies, 20(4): 1-15.
[8] Correia, R., Aksionova, E., Venciute, D., Sousa, J., Fontes, R. (2025). User-generated content's influence on tourist destination image: A generational perspective. Consumer Behavior in Tourism and Hospitality, 20(2): 167-185. https://doi.org/10.1108/CBTH-11-2023-0208
[9] Zhu, J., Zhan, L., Tan, J., Cheng, M. (2025). Tourism destination stereotypes and generative artificial intelligence generated images. Current Issues in Tourism, 28(17): 2721-2725. https://doi.org/10.1080/13683500.2024.2381250
[10] Miao, L., Yang, F.X. (2023). Text-to-image AI tools and tourism experiences. Annals of Tourism Research, 102: 103642. https://doi.org/10.1016/j.annals.2023.103642
[11] Zhang, J., Huang, J., Jin, S., Lu, S. (2024). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5625-5644. https://doi.org/10.1109/TPAMI.2024.3369699
[12] Chen, F.L., Zhang, D.Z., Han, M.L., et al. (2023). VLP: A survey on vision-language pre-training. Machine Intelligence Research, 20(1): 38-56. https://doi.org/10.1007/S11633-022-1369-5
[13] Zhu, H., Li, L., Wu, J., Zhao, S., Ding, G., Shi, G. (2022). Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. IEEE Transactions on Cybernetics, 52(3): 1798-1811. https://doi.org/10.1109/TCYB.2020.2984670
[14] Zhu, H., Zhou, Y., Li, L., Li, Y., Guo, Y. (2023). Learning personalized image aesthetics from subjective and objective attributes. IEEE Transactions on Multimedia, 25: 179-190. https://doi.org/10.1109/TMM.2021.3123468
[15] Gao, J., Li, P., Chen, Z., Zhang, J. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5): 829-864. https://doi.org/10.1162/neco_a_01273
[16] Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. The Visual Computer, 38: 2939-2970. https://doi.org/10.1007/s00371-021-02166-7
[17] Xu, Y., Yu, W., Ghamisi, P., Kopp, M., Hochreiter, S. (2023). Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks. IEEE Transactions on Image Processing, 32: 5737-5750. https://doi.org/10.1109/TIP.2023.3323799
[18] Yuan, Z., Hao, C., Zhou, R., et al. (2023). Efficient and controllable remote sensing fake sample generation based on diffusion model. IEEE Transactions on Geoscience and Remote Sensing, 61: 5615012. https://doi.org/10.1109/TGRS.2023.3268331
[19] Tang, D., Cao, X., Hou, X., Jiang, Z., Liu, J., Meng, D. (2024). CRS-Diff: Controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-14. https://doi.org/10.1109/TGRS.2024.3453414
[20] Liu, Y., Yue, J., Xia, S., Ghamisi, P., Xie, W., Fang, L. (2024). Diffusion models meet remote sensing: Principles, methods, and perspectives. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-22. https://doi.org/10.1109/TGRS.2024.3464685
[21] Li, C., Zhang, Z., Wu, H., et al. (2024). AGIQA-3K: An open database for AI-generated image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 34(8): 6833-6846. https://doi.org/10.1109/TCSVT.2023.3319020
[22] Chen, J., An, J., Lyu, H., Kanan, C., Luo, J. (2024). Learning to evaluate the artness of AI-generated images. IEEE Transactions on Multimedia, 26: 10731-10740. https://doi.org/10.1109/TMM.2024.3410672
[23] Zhang, W., Li, D., Ma, C., Zhai, G., Yang, X., Ma, K. (2023). Continual learning for blind image quality assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 2864-2878. https://doi.org/10.1109/TPAMI.2022.3178874
[24] De Lange, M., Aljundi, R., Masana, M., et al. (2022). A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3366-3385. https://doi.org/10.1109/TPAMI.2021.3057446