Application of Artificial Intelligence Generated Content (AIGC) in Contemporary Chinese Painting Creation

Enbo Fan
College of Sculpture and Public Art, Hebei Academy of Fine Arts, Shijiazhuang 050700, Hebei, China
Corresponding author: Enbo Fan (e-mail: optimuswk@163.com)

ABSTRACT While the current AIGC (Artificial Intelligence Generated Content) generation of Chinese paintings can reproduce the forms of ink and wash, it generally lacks the spirit of brush and ink and aesthetic conception, resulting in results that remain at the level of visual imitation. To address the lack of artistic spirit in AIGC, this paper constructs a generation system integrating brush and ink essence. It begins by collecting and semantically annotating a dataset with brushstrokes , white space, and artistic elements. CLIP multimodal encoding establishes image-to-semantic mapping. An a priori knowledge-based brush and ink control module is embedded within the Stable Diffusion framework to constrain generation, enhancing control over brushstroke flow and ink gradation. Philosophical concepts like virtuality/reality and white space are embedded using knowledge graphs. Interactive artist feedback is incorporated for collaborative creation. Experiments show significant improvements over the baseline: SSIM of 0.87 and FID of 18.5, a 40.3% improvement in overall aesthetic score, and a 26.2% improvement in semantic consistency CLIPScore. This confirms that the method effectively enhances the clarity, style, authenticity, and traditional artistic expression of generated Chinese paintings, while also supporting controllable visual pattern generation.

INDEX TERMS aigc,chinese painting generation,diffusion model,multimodal encoding,computational visualization

I. INTRODUCTION

As one of the most representative art forms of Chinese culture, Chinese painting has developed a unique aesthetic paradigm over its long history; similarly, embroidery arts from China, with their historical depth, have established unique and complex aesthetic paradigms regarding color harmony, weave structures, and symbolic motif placement. Painters focus on capturing spirit through form, emphasizing the vividness of spirit and the use of brushstrokes with a natural flow, pursuing the philosophical ideas and humanistic artistic conception embodied in brush and ink. This aesthetic orientation, centered on "artistic conception," enables Chinese painting to transcend the reproduction of form in expressing nature and conveying a state of mind, moving towards an exploration of the relationship between heaven, earth, and humanity. The "spirit of brush and ink" encompasses the force, rhythm, and gradation of ink color, while the "aesthetic artistic conception" embodies the subjective emotions and philosophical concepts conveyed by the painting. However, against the backdrop of the rapid development of contemporary information technology, artificial intelligence-generated content (AIGC) has been applied to the field of artistic creation [1,2]. Generated images often lack a sense of spirit and liveliness, and the artistic conception is insufficiently constructed, failing to convey the philosophical ideal of "harmony between man and nature." This situation not only limits the practical value of AIGC in Chinese painting creation but also keeps the application of artificial intelligence in traditional art at a nascent stage [3,4]. Therefore, how to enable AIGC to move beyond superficial imitation to a deeper appreciation of aesthetics has become a pressing issue in the exploration of the digitization of contemporary Chinese painting.

In recent years, scholars at home and abroad have tried to combine generative models with artistic creation and have explored this in the field of ink painting. In the interdisciplinary research field of AIGC and artistic creation, related research has shown a trend of multi-dimensional exploration. Du et al. [5] explored the collaborative generation mechanism of distributed diffusion models in wireless networks, providing a theoretical basis for the distributed deployment of AIGC technology. Lin et al. [6] focused on the integration of AI and Chinese calligraphy, exploring the innovative path of traditional art forms in the digital age. These studies provide important theoretical support and technical references for this paper, but further breakthroughs are still needed in the deep integration of the spirit of brush and ink and aesthetic conception in Chinese painting creation.

In response to the above shortcomings, some researchers have tried to improve the model effect by applying semantic information. In the study of the formal genealogy of contemporary Chinese painting, Yang [7] systematically analyzed the formal evolution of contemporary landscape painting, providing an important reference for understanding the modern transformation of traditional brush and ink language. Fengyuan et al. [8] further explored the digital technology and information expression methods of Chinese artists in the context of globalization, revealing the profound impact of new technologies on painting creation. Wang [9] provided a new dimension for understanding the social and cultural connotations of Chinese painting by analyzing the representation of female images in modern Chinese art history. Therefore, it is necessary to construct a multi-level generation framework that takes into account visual features, semantic conception, and philosophical connotations, so that artificial intelligence can not only reproduce the form of ink and wash, but also simulate and extend the spirit of brush and ink [10,11]. Against this background, this paper proposes an AIGC generation system that integrates the spirit of brush and ink. By combining data annotation, multimodal encoding, diffusion model optimization, and knowledge guidance, it explores new paths for artificial intelligence in contemporary Chinese painting creation.

This research aims to overcome the lack of aesthetic content in current AI-generated image processing (AIGC) technology when generating Chinese paintings, and to construct a generative system that embodies the spirit of brush and ink and artistic conception. The "human-computer collaboration" referred to in this paper specifically refers to the system generating an initial brushstroke composition, which the artist then modifies, adjusts layers, and refines their artistic conception according to their own creative intentions, feeding this feedback back to the model for iterative optimization. The "contemporary Chinese painting creation" referred to in this paper emphasizes the creative transformation and re-contextualization of traditional ink painting language using AI technology in the digital age. Although this system uses traditional brush and ink techniques (such as texturing, ink color division, and blank space) as its core modeling object, its generated results can serve as creative materials or sources of inspiration for contemporary artists, supporting their secondary creation in composition, media integration (such as digital ink painting + moving images), or expression of social issues (such as ecology and urbanization). Therefore, this method focuses on the "contemporary digital activation of traditional language," rather than directly generating images with socio-political content. Future work can build upon this foundation by introducing a contemporary visual symbol library to expand the system's expressive boundaries. This paper not only technically enhances the model's ability to represent brushstroke features, but also establishes a connection between artificial intelligence and the philosophical aesthetics of Chinese painting. The anticipated results provide new theoretical value and practical examples for the application of artificial intelligence in traditional art, and open up innovative pathways for the digital transformation of contemporary Chinese painting.

II. GENERATIVE DESIGN THAT INTEGRATES THE SPIRIT OF BRUSH AND INK

A. SEMANTIC ANNOTATION AND DATASET CONSTRUCTION

The dataset for this study is collected from three sources: first, high-definition digitized Chinese paintings collected in public art databases; second, digital resource libraries provided by museums and art schools; and third, manually scanned paper albums and scrolls. During data collection, representative works of landscape, flower-and-bird, and figure painting are prioritized to ensure coverage of the main themes of Chinese painting. To reduce interference from noisy samples, the collected images (Figure 1) are resized to a uniform resolution of 512×512 pixels, and a bilateral filter is applied to remove artifacts generated during the scanning process. A preliminary sample of 18,642 images is collected, including 8,203 landscapes, 6,421 flower-andbird, and 4,018 figures. To ensure data balance, random downsampling is used to limit the sample size to 15,000 images, maintaining a near 4:3:3 ratio among the three themes in the training set.

The initial sample consists of 18,642 images. The final semantically labeled dataset contains 14,263 images with a total of 171,156 labels, averaging 11.99 labels per image. To ensure the reliability of the annotations, all semantic labels are independently annotated by three graduate students with a background in Chinese painting. The Fleiss' Kappa coefficient (κ=0.78) is then calculated, and samples with a consistency score below 0.7 undergo arbitration for verification. Furthermore, the annotator consistency matrix shows that the consistency of most label pairs is between 0.7 and 0.9, confirming the high reliability of the annotation process. The annotation process is completed using LabelMe and a self-developed Python script. For brush stroke areas, an automatic pre-annotation method based on SLIC (Simple Linear Iterative Clustering) superpixel segmentation is used to divide the image into several uniform regions and calculate the grayscale variance of each region to assist in determining the ink color density. Manual annotators make corrections based on this to improve efficiency. The white space ratio is calculated through binarization and regional connectivity analysis, as shown in Formula (1):

FIGURE 1. Chinese paintings in the database

$$L_r = \frac{A_w}{A_t} \tag{1}$$

L_r represents the white space ratio; A_w represents the number of white pixels in the image; A_t represents the total number of pixels. The calculated results are used to annotate the white space distribution and stored in a JSON file format. For artistic conception elements, a text description mapping method is used to convert manually annotated keywords into vector representations using Word2Vec, ensuring that they correspond to image feature vectors during subsequent model training.

After annotation, the data is uniformly stored in the COCO (Common Objects in Context) format, which includes image paths, annotated region coordinates, and semantic label information. To ensure the quality of the training data, the dataset undergoes a two-step quality control process: first, a cross-check is performed on 5% of the samples, removing samples with an inconsistency rate exceeding 20%. Second, statistics are performed on the entire data set to confirm that the distribution of the number of stroke labels, the white space distribution, and the distribution of artistic conception keywords are consistent with the subject matter. The resulting semantically annotated dataset consists of 14,263 images with a total of 171,156 labels, for an average of 11.99 labels per image.

Figure 2 shows the results of constructing a painting semantic dataset for the context of new architectural industrialization. Figure 2(a) displays the semantic feature distribution of works of different subject matter in a three-dimensional coordinate space. The x-axis represents brushstroke complexity; the y-axis represents ink color depth; the z-axis represents white space ratio. These correspond to the semantic feature distributions of landscape painting, flower-and-bird painting, and figure painting, respectively. It can be seen that the three types of paintings exhibit a clear clustering trend in the three-dimensional semantic space. For example, landscape painting exhibits higher brushstroke complexity and ink density, while flower-and-bird painting is more prominent in white space ratio and ink color depth. Figure painting is concentrated in the white space ratio dimension, reflecting the differences in the expressive styles of the three types of subject matter. Figure 2(b) shows the annotator consistency matrix. Colors closer to red indicate higher consistency, with an overall range of 0.7–0.9, indicating that most annotators reach a high degree of consensus on semantic label judgment, providing assurance for the reliability of the dataset.

B. MULTIMODAL ENCODING OF INK PAINTING STYLES

1) Image and Semantic Feature Extraction

Based on a semantically annotated dataset, it is necessary to establish a correspondence between images and semantics. To this end, CLIP (Contrastive Language-Image Pretraining) is used as a multimodal encoding framework. This model consists of two parts: an image encoder and a text encoder. The image encoder uses the Vision Transformer (ViT-B/32) architecture, with an output feature dimension of 512. All images are uniformly normalized before being input to the encoder and center-cropped to minimize edge noise. The text encoder uses a Transformer architecture, with input consisting of a sequence of segmented semantic labels, whose vector dimensions match the image features. This design ensures that image and semantic features are aligned in the same vector space.

During the training process, the contrastive learning loss function is used to optimize the correspondence between images and semantics. Its goal is to maximize the similarity between matching samples and minimize the similarity between mismatching samples [12]. The loss function is as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\mathrm{sim}(v_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(v_i,t_j)/\tau)} + \log\frac{\exp(\mathrm{sim}(t_i,v_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(t_i,v_j)/\tau)}\right] \tag{2}$$

v_i represents the i-th image feature; t_i represents the corresponding semantic label feature; $\mathrm{sim}(\cdot)$ represents the cosine similarity function; τ represents the temperature parameter, set to 0.07. This study trains the model for 30 epochs on four A100 GPUs (Graphics Processing Units), with a batch size of 256 and a learning rate schedule using cosine annealing, with an initial value of e^{-4}.

C. ALIGNMENT BETWEEN BRUSHWORK FEATURES AND ARTISTIC CONCEPTION

During the feature alignment phase, the image's brushstroke characteristics, ink color levels, and white space ratio are

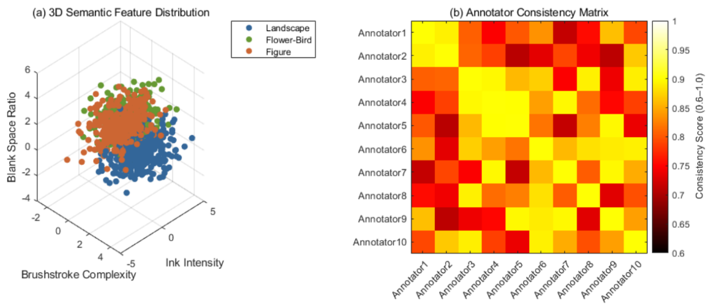

FIGURE 2. Visualization of semantic annotation and dataset construction. (a) Three-dimensional distribution of semantic features of three categories of topics; (b) Consistency matrix of annotators

input into the CLIP image encoder as part of the image feature vector. Semantic labels are derived from the dataset annotations from the previous phase, including texturing techniques, ink color density, white space ratio, and keywords related to artistic conception. To ensure that the semantic vectors accurately reflect the differences between labels, Word2Vec pre-trained word vectors are used for the keywords, with a dimension of 300. These are then mapped to the same 512-dimensional space as the image features through a fully connected layer.

During the training process, the correspondence between image features and semantic features is continuously optimized through contrastive learning. For identical image and semantic pairs, the model strengthens their proximity in the vector space; for mismatched samples, the model increases the distance between them [13]. In this way, the model gradually learns the correspondence mechanism between “brush and ink features and semantic imagery”. After training, when a semantic description is input, the model retrieves or generates matching image features in the feature space, providing semantic constraints for the conditional generation of the subsequent diffusion model.

Figure 3 shows the alignment quality distribution of image and text feature intensities for three themes. The horizontal axis represents image feature intensity, and the vertical axis represents text feature intensity, with the height of the vertical axis representing alignment quality. Landscape themes peak at a feature intensity close to 0.5, indicating that moderate-intensity features are easier to align. Flower and bird themes exhibit peaks at 0.4 for image features and 0.6 for text features, demonstrating certain structural differences. Figure themes reach their maximum values at 0.6 for image features and 0.4 for text features, indicating higher cross-modal alignment within this range. Surface plot visualization verifies the model’s ability to differentiate semantic alignment between image and text across different themes.

D. FEATURE SPACE CONSTRUCTION AND REGULARIZATION CONSTRAINTS

During the training process, in order to avoid feature space collapse and overfitting, a variety of regularization methods are used. First, random occlusion and color perturbation are applied in the image input stage to ensure the robustness of the model to changes in ink color. Second, Dropout is added to the semantic vector encoding with a ratio of 0.1 to prevent the model from over-relying on certain keywords. Finally, orthogonal constraints are applied in the feature space construction stage. By adding regularization terms to the training loss, the vectors of different semantic categories remain independent. The regularization term is in the form of Formula (3) [14]:

$$R = \lambda \sum_{i \neq j} \left(t_i^{\top} t_j \right)^2 \tag{3}$$

t_i and t_j represent semantic vectors of different categories, and λ is a weight parameter set to 0.01. This mechanism effectively avoids confusion in the semantic space, thereby ensuring the distinction between the brushwork and artistic conception labels.

E. OPTIMIZATION OF INK GENERATION FOR DIFFUSION MODEL

The "prior knowledge" referred to in this paper refers to quantifiable visual constraints statistically derived from a large number of traditional Chinese paintings. Specifically, it manifests as two sets of soft regularization mechanisms: first, a priori knowledge of brushstroke continuity based on optical flow fields, which dynamically constrains brushstroke flow by calculating the consistency of pixel motion between adjacent denoising steps; second, a priori knowledge of ink color distribution based on gray-level histogram matching. This priori knowledge originates from the statistical distribution of ink color levels (256-level gray-level histogram) of real ink paintings in the training set, and uses Kullback-Leibler divergence as a loss term to guide the ink color gradation of the generated image to approximate the traditional paradigm. These two types of priors are not rigid rules, but are embedded in the denoising objective of the stable diffusion model in the form of differentiable loss functions. This ensures that while preserving the diversity of generation, the output conforms to the visual characteristics of traditional Chinese painting in terms of brushstroke dynamics and ink color levels.

After establishing the multimodal feature space, Stable Diffusion is used as the basic generative framework for ink

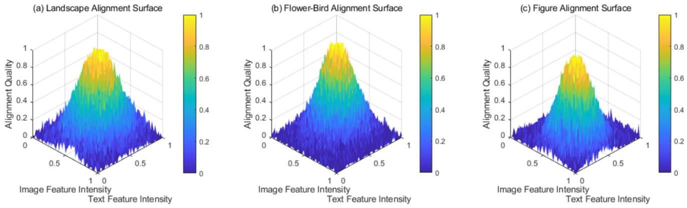

FIGURE 3. Multimodal alignment quality surfaces. (a) Landscape-themed image-text alignment quality surface; (b) Flower-and-bird-themed image-text alignment quality surface; (c) Figure-themed image-text alignment quality surface

painting style optimization. The model consists of a forward diffusion process and a backward generation process. The forward diffusion process converts the real image into pure noise by gradually adding Gaussian noise, while the backward generation process uses a neural network to gradually denoise and restore the target image. To adapt to the semantic and brushwork characteristics of Chinese painting, the main UNet network is restructured, replacing the original residual blocks with a module combining 2D convolution and attention mechanisms to enhance the modeling of local texture and long-term image dependencies. The training input consists of an image-semantic feature vector aligned by CLIP. The image resolution is fixed at 512×512; the time step is set to 1000; the noise scheduling uses a linear β schedule ranging from $\beta_1 = 1 \times 10^{-4}$ to $\beta_T = 0.02$.

The traditional diffusion model relies solely on convolutional features to restore image structure during denoising, making it difficult to express the unique dynamic brushstroke effects of Chinese painting. To address this problem, a brushstroke fluidity module based on optical flow constraints is applied. This module estimates the optical flow field F_t on the intermediate image generated at each time step and calculates the brushstroke consistency loss between adjacent time steps, as shown in Formula (4):

$$L_{\text{flow}} = \frac{1}{N} \sum_{i=1}^{N} \left\| x_t^i - \text{Warp}\left(x_{t-1}^i, F_t\right) \right\|_2^2 \tag{4}$$

x_t^i represents the generated result for the i-th pixel at time step t, and $\text{Warp}(\cdot)$ is the optical flow field transformation function. This loss term helps the model maintain the continuity and fluidity of brushstrokes during iteration, avoiding ink breaks and local discontinuities. This mechanism ensures that the generated images more closely resemble the dynamic characteristics of hand-painted strokes in terms of brushstroke motion. To enhance the layering of ink rendering, an ink color constraint module based on histogram matching is applied in the reverse generation stage. First, the ink color distribution in the training set is statistically calculated as a histogram H_{ref}, which is divided into 256 grayscale levels. During the generation process, the grayscale histogram H_{gen} of the current output is calculated at each time step, and the Kullback-Leibler divergence is applied as a regularization term [15], as shown in Formula (5):

$$L_{\text{ink}} = \sum_{k=1}^{256} H_{\text{ref}}(k) \log \frac{H_{\text{ref}}(k)}{H_{\text{gen}}(k)} \tag{5}$$

By minimizing this loss, the model gradually adjusts the ink color distribution of the output image to make it closer to the light and dark levels of a real ink painting. At the same time, a channel attention mechanism is applied at the feature level to give higher weights to dark areas, ensuring that the levels in the shadows and thick ink parts are clear. In the experimental setting, the regularization weight parameter is set to $L_{\text{ink}} = 0.05$ to avoid excessively suppressing the degree of freedom of generation while ensuring style constraints. The final training objective is composed of the reconstruction loss, the flow loss, and the ink level loss, as shown in Formula (6):

$$L_{\text{total}} = L_{\text{recon}} + \alpha L_{\text{flow}} + \beta L_{\text{ink}} \tag{6}$$

L_{recon} represents the mean squared error reconstruction loss of the standard diffusion model, while α and β are weight parameters set to 0.1 and 0.05, respectively. This combined loss ensures the generated image's fidelity while enhancing the dynamics of brushstrokes and the gradation of ink. During the optimization process, mixed-precision training is used to reduce video memory usage, and gradient clipping is applied in the later stages of training with a threshold of 1.0 to prevent gradient explosion.

Figure 4 shows the evolution of the feature distribution of the diffusion model at different time steps, with the eigenvalue on the horizontal axis and the probability density on the vertical axis. It can be seen that in the early iteration (t=10), the distribution is relatively dispersed and has a large variance, indicating that the features at this stage are still close to random noise. As the iterations advance to t=100 and t=500, the distribution gradually shrinks and forms a clear peak, indicating that the model begins to capture stable semantic features. At the late iteration (t=900), the distribution is highly concentrated near the mean, indicating that the model is ultimately able to generate feature representations with clear structure and stable semantics. This result intuitively demonstrates the gradual convergence of the diffusion model from noise to clear ink-and-wash semantics, and strongly demonstrates the role of optimization strategies in improving generation quality.

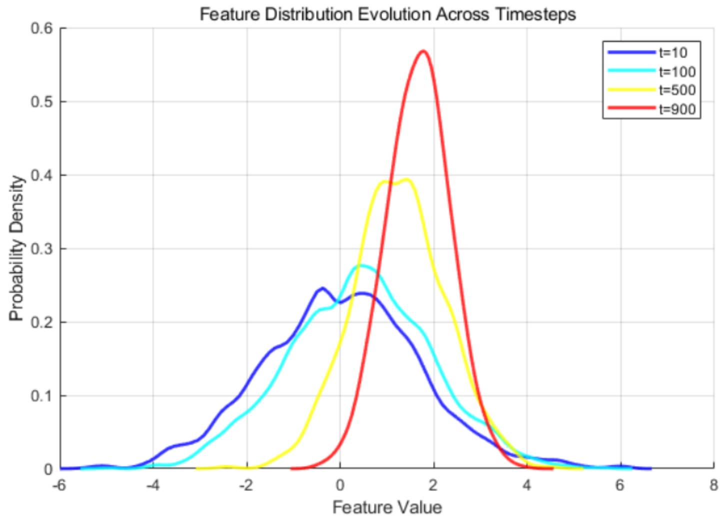

FIGURE 4. Evolution of feature distribution at different time steps

F. KNOWLEDGE GUIDANCE ON AESTHETIC CONCEPTION

To formalize the core compositional principles of Chinese painting, this study operationalizes the "relationship between emptiness and fullness" in traditional aesthetics as "the proportional and interactive relationship between non-physical areas (blank space) and physical brushstroke areas in spatial layout." "Blank space layout" is quantified as the number and area proportion of connected white areas in the image, and their distribution entropy in the foreground-middleground-background three-layer space. "The unity of man and nature" is modeled through element co-occurrence constraints, such as the high-frequency co-occurrence pattern of the "figure-landscape" combination under specific compositional proportions. These operationalizations are encoded as structured nodes and edge weights in a knowledge graph, thus providing computable layout guidance during the generation process.

First, artistic conception keywords and their corresponding image features are extracted from the semantic annotation dataset to form a node set. Nodes include pictorial elements (such as rocks, trees, and figures), brushwork characteristics (such as texture type and ink color density), composition layout (such as the proportion of foreground, midground, and background), and artistic conception vocabulary (such as "virtual," "distant," and "quiet"). Edge relationships are defined as logical or spatial associations between concepts, such as "white space-virtual-real relationship" and "rocks-perspective perspective." The graph is stored in an RDF (Resource Description Framework) structure. Each edge contains relationship type and weight information. The weight is calculated based on node co-occurrence frequency and artist annotation experience, and is set in the range of 0.0–1.0 [16,17]. Through graph construction, the model obtains constraint information between nodes during generation, achieving the encoding of the aesthetic structure of Chinese painting.

Figure 5 above illustrates the feature structure and consistency assessment guided by aesthetic conception knowledge. Figure 5(a) shows a 3D knowledge graph. Each node in the three-dimensional space represents a concept, including aesthetic concepts such as "virtual-real relationship," "white space layout," and "natural harmony," as well as their specific elements such as "stone," "tree," and "water flow." The thickness of the edges corresponds to the strength of the association, revealing the hierarchical and semantic connections between aesthetic concepts and compositional elements in ink painting. Figure 5(b) shows a knowledge consistency loss surface. The X-axis and Y-axis represent knowledge dimension 1 and knowledge dimension 2, respectively, and the Z-axis represents the loss value for different combinations of knowledge dimensions. The surface shows that when the model's representations of different knowledge dimensions are inconsistent, the loss increases, while the loss is lower in the range of conceptual fit, demonstrating the important role of knowledge guidance in maintaining semantic and aesthetic consistency. Figure 5 intuitively illustrates how knowledge modeling can be used to constrain the generative model, allowing it to better align with the aesthetic conception of ink painting during the creative process.

In the generation phase, the knowledge graph is embedded in the diffusion model as a conditional constraint. First, the knowledge graph nodes are vectorized using a graph convolutional network (GCN). Assuming that the dimension of each node is 128, after two layers of convolution and ReLU activation, the final semantic feature matrix $K \in R^{N\times128}$ is output, where N is the number of nodes, with an average of 512 matrices. Subsequently, K is fused with the image-semantic features encoded by CLIP through the attention mechanism [18], and the attention weight matrix is calculated, as shown in Formula (7):

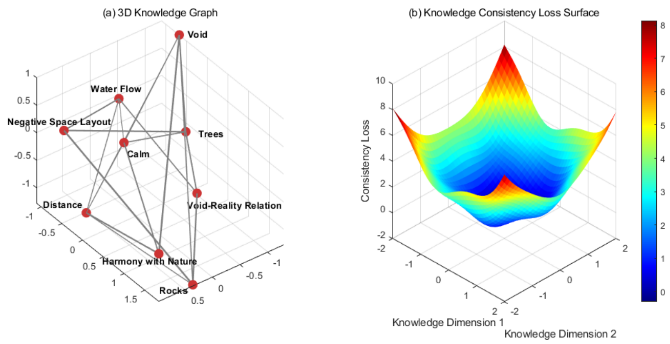

FIGURE 5. 3D knowledge graph and consistency loss graph. (a) 3D knowledge graph; (b) Knowledge consistency loss surface

$$A = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{7}$$

Q is the image-semantic feature matrix; K is the knowledge graph feature matrix; $d_k = 128$ is the feature dimension; A represents the knowledge guidance weight matrix. Through this mechanism, the model references the nodes and relationships in the knowledge graph during the denoising process, ensuring that the generated brushstroke layout and artistic conception characteristics conform to the philosophical structure of Chinese painting. To ensure that knowledge guidance plays a role in the generation process, the knowledge consistency loss L_{kg} is defined as follows:

$$L_{\mathrm{kg}} = \frac{1}{N}\sum_{i=1}^{N} \|f_{\mathrm{gen}}(x)_i - f_{\mathrm{kg}}(K)_i\|_2^2 \tag{8}$$

$f_{\mathrm{gen}}(x)_i$ is the feature vector corresponding to node i in the generated image, and $f_{\mathrm{kg}}(K)_i$ is the feature vector of the knowledge graph. The loss function constrains the similarity between the generated features and the knowledge graph features using mean squared error, ensuring that the virtual-real relationship, white space layout, and spatial hierarchy are reflected in the generated image. The weight parameter λ_{kg} is set to 0.08 to balance the contribution of brushstroke flow and ink gradation loss.

During the back-diffusion process, the generator uses the conditional vector provided by the knowledge graph at each denoising step to adjust pixel distribution and local texture through attention weighting. Specifically, each attention block in the UNet applies knowledge-conditioned input to ensure that local region generation matches the knowledge constraints. The total step size for iterative training is 1000, and the noise scheduling is consistent with the diffusion model described above. The comprehensive loss is optimized through gradient descent [19,20]. This strategy continuously constrains the image layout and artistic conception at each generation step, ensuring that the model output aligns with the characteristics of traditional Chinese painting in both form and philosophical structure.

G. ARTISTIC REVISION OF HUMAN-MACHINE COLLABORATION

The interactive artist feedback mechanism in this system primarily guides local searches in the latent space, rather than directly optimizing model weights. Specifically, when an artist modifies brushstrokes or adjusts ink color in a local area of the generated image on the WebGL frontend, the system transforms this operation into an incremental vector Δz in the latent space through inverse mapping, and uses it as a conditional input superimposed on the latent representation of the current denoising step. This Δz is extracted from the user's operation trajectory by a lightweight encoder and continuously applied in subsequent denoising steps 1–3 to ensure that the intention to modify is preserved in the generated result. The entire process does not trigger model retraining, but rather achieves real-time collaboration at the single-sample level through conditional guidance, thereby supporting the artist's intervention in local aesthetic details while maintaining global style consistency.

To achieve collaborative creation between AI and artists, this study designs and implements an interactive feedback system. The system is built on WebGL and PyTorch backends, supporting real-time display and manipulation of high-resolution images. Painters can modify local brushstrokes, adjust ink density, and adjust white space ratio on the generated image, while also supporting zooming and local area selection. Each operation is converted into a potential feature change vector by the system and used as a conditional input to guide the model to follow the artist's modification intention in subsequent generation iterations [21]. This mechanism allows the painter to directly intervene in the generation process, ensuring that the generated image conforms to the aesthetic core of Chinese painting and enhancing the artistic value of the work. After interactive modification, the system fine-tunes the intermediate features of the generated model. Local incremental feature vectors are applied to the UNet encoding layer to reflect the artist's adjustments to brushstrokes and ink color. The magnitude of fine-tuning is controlled to ensure global stylistic consistency. Fine-tuning operations are combined with batch iterations, allowing multiple local adjustments per image. The system rapidly updates the generated image with each iteration to reflect the artist's aesthetic judgment. During training, regularization constraints are used to maintain consistency between local modifications and the overall style, ensuring that local adjustments do not disrupt the fluidity of brushstrokes and the gradation of ink colors, while maintaining the overall composition and artistic conception of the generated image. To further ensure that the generated image retains the characteristics of Chinese painting after local modifications, this study applies multiple constraint mechanisms to the generation process. First, by diffusing the model's existing brushstroke flow and ink color layer optimization modules, it ensures that local modifications do not disrupt the overall dynamics and layering. Second, by embedding the aesthetic constraints of the knowledge graph into the attention mechanism, the generated image maintains the coordination of virtual-real relationships, white space layout, and overall artistic conception while undergoing local modifications.

III. TEST OF THE EFFECT OF INTEGRATING THE SPIRIT OF BRUSH AND INK

The dataset used in the evaluation experiments of this study is expanded from a previously constructed dataset of Chinese paintings that incorporates the spirit of brush and ink, covering three categories of subject matter: landscapes, flowers and birds, and figures. The dataset also contains manually annotated brushstroke features, ink color levels, white space layout, and artistic conception keywords, providing a basis for the semantic consistency evaluation of the generated results. To ensure the fairness of the evaluation, the samples in the dataset are randomly shuffled and assigned generation tasks to ensure a balanced representation of each category of subject matter. Some samples are used as reference images for the calculation of visual and style indicators, while the remaining samples are used for expert scoring and semantic matching measurement.

A. IMAGE CLARITY ASSESSMENT

To measure the clarity of generated images, the Structural Similarity Index (SSIM) is used for evaluation. In the experiment, the generated images are paired with corresponding high-quality, real-world Chinese paintings, and a structural similarity score is calculated to reflect the image consistency in brightness, contrast, and texture. All images are resized to 512×512 pixels and normalized to grayscale before calculation. During the evaluation process, each image pair is scanned multiple times and segmented into local regions. The average global and local SSIM scores are calculated to obtain a comprehensive representation of the generated image's detail fidelity and overall visual structure. This method can intuitively assess a model's ability to reproduce ink textures, brushstroke edges, and ink color transitions. Evaluation methods include: Proposed, SD Baseline (Stable Diffusion), GAN-based (Generative Adversarial Network), VAE-based (Variational Autoencoder), Style Transfer, CNN-based (Convolutional Neural Network), Rule-based, and Hybrid, as shown in Figure 6 .

Figure 6 effectively demonstrates the proposed method's superiority in preserving image structure and detail clarity, quantified by the SSIM value. Figure 6(a) shows the proposed method achieves an SSIM of 0.87, significantly outperforming all baseline methods. Figure 6(b) further analyzes clarity across subjects (landscape, flowers/birds, people), confirming the stable performance of the proposed method. For instance, landscape subjects reach a maximum SSIM of 0.89, and even the lowest-performing subject, people, achieves 0.84, both substantially exceeding the SD Baseline's range of 0.70-0.75. Overall, these results confirm the method's effectiveness in maintaining high image detail and structural fidelity across varied artistic content.

B. STYLE FIDELITY EVALUATION

To assess the stylistic similarity between generated images and real-world Chinese paintings, the Fréchet Inception Distance (FID) metric is used. First, the generated and real-world images are fed into a pre-trained Inception-V3 model to extract deep feature representations. The mean and covariance differences between the two feature distributions are then calculated to quantify the degree of deviation in the generated results in the style space. To ensure accurate evaluation, the images are augmented using random rotation, cropping, and color normalization to reduce the influence of incidental noise. This metric objectively reflects the model's overall performance on stylistic elements such as brushstrokes, ink gradation, and white space layout, facilitating horizontal comparison with other generative methods.

Figure 7 demonstrates the proposed method's superior style fidelity, evidenced by the low FID (Fréchet Inception Distance). Figure 7(a) compares eight methods, showing the proposed approach achieved the lowest FID of 18.5, indicating its generated images are stylistically the closest to the real data. Figure 7(b) illustrates the convergence behavior: the proposed method's FID curve drops rapidly and stabilizes around 19, demonstrating better optimization efficiency and style consistency compared to the slower-converging SD Baseline. Furthermore, Figure 7(c), which plots the 2D feature space distribution, shows that the feature cluster (red dots) for the generated images closely aligns with the real image features (green dots). This visual analysis confirms that the proposed method effectively maintains high style consistency within the multidimensional feature space.

C. AESTHETIC EXPERT RATING

To evaluate the generated images from an artistic aesthetic perspective, an expert panel consisting of five professional Chinese painters and three art scholars is invited to evaluate the generated images. The experts rate the generated images in a unified environment, assessing each dimension on a scale of 1–10, including vividness of spirit and charm, brushwork, artistic conception, white space, and ink gra-

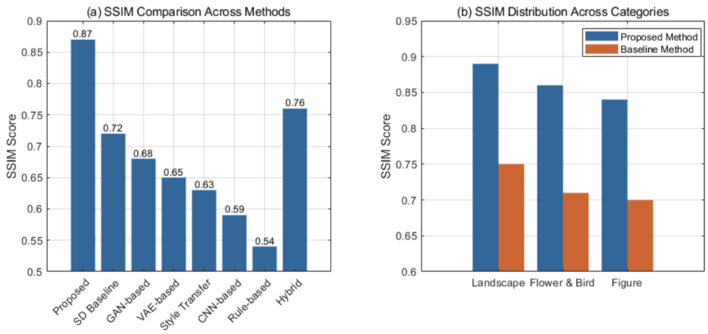

FIGURE 6. Image clarity evaluation. (a) Comparison of SSIM values for different methods; (b) Distribution of SSIM values for different themes

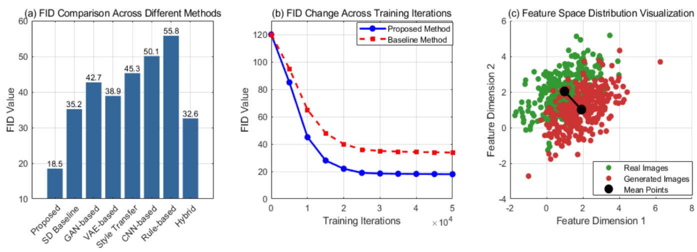

FIGURE 7. Style fidelity evaluation. (a) Comparison of FID values of different methods; (b) FID value changes with training iterations; (c) Visualization of feature space distribution

dation. Prior to scoring, all images are randomly shuffled to avoid order bias. During the scoring process, the experts are asked to observe each image from both an overall perspective and detailed observations, and to record their suggestions for revisions. Statistical analysis is then performed to determine the average scores for each dimension and a comprehensive score, reflecting the artistic quality of the generated images according to traditional aesthetic standards.

In the aesthetic evaluation of Chinese paintings generated by artificial intelligence, the data in Table 1 shows that the proposed method is significantly better than the baseline method in all traditional aesthetic dimensions. Among them, the artistic conception expression dimension has the most significant improvement, from 5.8 points to 8.9 points, an increase of 53.4%. The comprehensive score reaches 8.7 points, an increase of 40.3% compared with the baseline method's 6.2 points, indicating that the model has approached professional level in core aesthetic elements such as vivid spirit and bone-like brushwork.

TABLE 1. Aesthetic expert scoring data

Evaluation Dimension	Proposed Method Score	Baseline Method Score	Improvement
Vitality and Spirit	8.8	6	46.7%
Brushwork Technique	8.5	6.3	34.9%
Artistic Conception	8.9	5.8	53.4%
Blank Space Layout	8.6	6.5	32.3%
Ink Tone Hierarchy	8.7	6.4	35.9%
Overall Score	8.7	6.2	40.3%

D. SEMANTIC CONSISTENCY ASSESSMENT

To verify the degree of match between the generated images and the input semantic labels, it uses CLIPScore to perform semantic consistency assessment. The experimental steps involve inputting the generated images and corresponding semantic labels into the CLIP model, extracting feature vectors for the image and text, and calculating cosine similarity. A batch calculation is performed on the entire dataset, and the mean and standard deviation of the consistency scores between the generated images and the semantic labels are calculated to quantify the accuracy of the generated results in terms of semantic expression. This method verifies whether the model accurately understands the brushstroke characteristics, white space distribution, and artistic conception keywords in the input semantics, thereby determining whether the generated images conform to the expected semantic content.

Table 2 demonstrates the proposed method's significant superiority in semantic consistency (CLIPScore) across several subcategories compared to the baseline. The proposed method consistently achieves higher scores and lower standard deviations (indicating greater stability). For key elements, the proposed method achieves scores of 0.85±0.03 for brushstroke features (a 25.0% improvement over the baseline's 0.68±0.05); 0.83±0.04 for ink gradation (a 27.7% improvement); 0.81±0.05 for white space distribution, which sees the highest improvement (28.6%). The overall average score for the proposed method is 0.82±0.04, reflecting an average relative consistency improvement of 26.2% over the baseline's 0.65±0.06. These results confirm that the method substantially improves the model's ability to capture unique semantic elements of ink painting, especially spatial composition and ink color expression.

TABLE 2. Semantic consistency evaluation data table

Semantic Category	Proposed Method CLIPScore	Baseline Method Score	Consistency Improvement
Brushwork Features	0.85±0.03	0.68±0.05	25.0%
Ink Tone Levels	0.83±0.04	0.65±0.06	27.7%
Blank Space Distribution	0.81±0.05	0.63±0.07	28.6%
Artistic Concept Keywords	0.79±0.06	0.62±0.08	27.4%
Overall Semantics	0.82±0.04	0.65±0.06	26.2%

E. DETAIL FIDELITY EVALUATION

To further evaluate the detailed representation of brushstroke textures and ink color gradations in generated images, a local feature comparison metric is applied. The specific method involves selecting key local regions in the generated and real images, such as rock textures, tree branches, and details of clothing, and using highresolution convolutional features for feature extraction and similarity calculation. By comparing and analyzing different regions, a local detail fidelity score is obtained, and the average score for each subject is calculated. This metric directly reflects the model's ability to reproduce the microscopic brushwork features and local structure of Chinese paintings, providing a more refined evaluation dimension for generation quality. Figure 8 shows the evaluation results of the generated images in terms of detail fidelity. The y-axis in Figure 8(a) represents the detail fidelity score. The proposed method (Proposed) performs best in detail fidelity, with a score of 0.88, higher than other methods such as the SD Baseline (0.72) and the GAN-based method (0.65). The traditional Rule-based method has the lowest score, at only 0.54, indicating that the model is superior in restoring texture and local details. Figure 8(b) shows the proposed method selecting key local regions such as people, rocks, and leaves in a real image, demonstrating that the model has good results in restoring detailed areas.

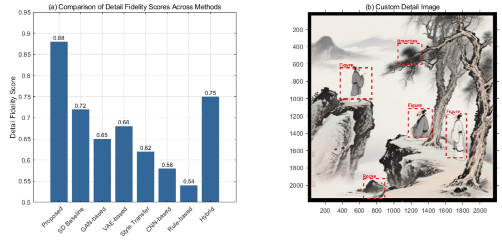

FIGURE 8. Detail fidelity evaluation. (a) Comparison of detail fidelity of different methods; (b) Image display of detail area

IV. CONCLUSION

This paper, focusing on the application of Artificial Intelligence Generated Content (AIGC) in contemporary Chinese painting, proposes a multimodal generation system that integrates the spirit of brush and ink with aesthetic conception.This study constructs a semantically annotated dataset, employs CLIP multimodal encoding to establish a correspondence between brush and ink features and the semantics of artistic conception, optimizes the brushstroke flow and ink color gradation of a stable diffusion model, and embeds philosophical concepts through knowledge graphs to achieve human-machine collaborative correction of the generation process. Experimental results demonstrate that this method performs well across multiple evaluations. First, the overall aesthetic score improves by 40.3%, bringing the style closer to authentic ink paintings. Regarding semantic consistency, the CLIPScore improves by 26.2%, indicating that the generated images are more consistent with the input semantics. In terms of objective metrics, the SSIM reaches 0.87, and the FID decreases to 18.5, demonstrating an overall improvement in image clarity and quality. Overall, the results demonstrate that AIGC significantly enhances the stylistic authenticity, clarity, and artistic expression of generated images, effectively reproducing the form and artistic conception of Chinese painting.

AUTHOR CONTRIBUTIONS

Enbo Fan designed, collected and analyzed the data, and drafted the manuscript. Enbo Fan conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Enbo Fan participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately inve stigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Guo, Y. Lu, Y. Dou, and F. Wang, "Can ChatGPT boost artistic creation: The need of imaginative intelligence for parallel art," IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 4, pp. 835-838, 2023, doi: 10.1109/JAS.2023.123555.

[2] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," Artificial intelligence review, vol. 55, no. 1, pp. 589-656, 2022, doi: 10.1007/s10462-021-10039-7.

[3] T. E. Onyejelem and E. M. Aondover, "Digital generative multimedia tool theory (DGMTT): A theoretical postulation in the era of artificial intelligence," Adv Mach Lear Art Inte, vol. 5, no. 2, pp. 01-09, 2024, doi: 10.17265/2160-6579/2024.03.004.

[4] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "Generative artificial intelligence: a systematic review and applications," Multimedia Tools and Applications, vol. 84, no. 21, pp. 23661-23700, 2025, doi: 10.1007/s11042-024-20016-1.

[5] H. Du, R. Zhang, D. Niyato, J. Kang, Z. Xiong, D. Kim, et al., "Exploring collaborative distributed diffusion-based AI-generated content (AIGC) in wireless networks," Ieee network, vol. 38, no. 3, pp. 178-186, 2023, doi: 10.1109/MNET.006.2300223.

[6] T. T. Y. Lin, J. She, Y. A. Wang, and K. Zhang, "Future ink: The collision of AI and Chinese calligraphy," ACM Journal on Computing and Cultural Heritage, vol. 18, no. 1, pp. 1-17, 2025, doi: https:10.1145/3700882.

[7] Q. Yang, "Contemporary Chinese landscape paintings: genealogies of form," Journal of Visual Art Practice, vol. 22, no. 4, pp. 347-371, 2023, doi: 10.1080/14702029.2023.2264076.

[8] Z. Fengyuan, D. Wongsarot, and S. Rohitasuk, "Digital Technology and Information Expression in the Painting Works of Chinese Artists Born in the 1990s in the Era of Globalization," Journal of Multidisciplinary in Humanities and Social Sciences, vol. 7, no. 3, pp. 1256-1271, 2024, doi: https://so04.tci-thaijo.org/index.php/jmhs1_s.

[9] S. Wang, "Fashioning Chinese feminism: Representations of women in the art history of modern China," Critical Studies in Fashion & Beauty, vol. 12, no. 2, pp. 207-227, 2021, doi: 10.1386/csfb_00027_1.

[10] C. Ma, "Research on Design Realization and Dissemination in Traditional Painting Appropriation in the Context of Digital Technology," Advances in Humanities and Modern Education Research, vol. 1, no. 1, pp. 91-97, 2024, doi: 10.70114/ahmer.2024.1.1.P91.

[11] C. Ribeiro, "Chinese contemporary art: where it comes from, where it goes," Asiadémica: revista universitaria de estudios sobre Asia Oriental, no. 17, pp. 241-254, 2022.

[12] R. Li, "From mechanism to organicism: thoughts on linear perspective and re-contextualizing contemporary Chinese art," World Art, vol. 12, no. 2, pp. 195-211, 2022, doi: 10.1080/21500894.2022.2066167.

[13] Z. Lei and W. Zhou, "Reinterpretation of the autonomy of art in the contemporary environment," International Communication of Chinese Culture, vol. 10, no. 2, pp. 203-221, 2023, doi: 10.1007/s40636-023-00277-5.

[14] X. Lin, "Protecting AI-generated images as works of fine art in China: Magnifying the legacy of art to copyright," Law, Technology and Humans, vol. 7, no. 1, pp. 107-126, 2025, doi: 10.5204/lthj.3825.

[15] X. Duan, "A study of Chinese and Western landscape painting from the comparative perspective," Frontiers in Art Research, vol. 5, no. 13, pp. 1-7, 2023, doi: 10.25236/FAR.2023.051301.

[16] X. Chen, Q. Liu, Y. Chen, R. Wang, Y. You, and W. Deng, "ColorNetVis: An Interactive Color Network Analysis System for Exploring the Color Composition of Traditional Chinese Painting," IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 6, pp. 2916-2928, 2024, doi: 10.1109/TVCG.2024.3388520.

[17] W. Zhang, D. Ren, and G. Legrady, "Cangjie's poetry: an interactive art experience of a semantic human-machine reality," Proceedings of the ACM on Computer Graphics and Interactive Techniques, vol. 4, no. 2, pp. 1-9, 2021, doi: 10.1145/3465619.

[18] S. Yanan, A. Kamis, and H. Pratama, "Current approaches and future trends in integrating digital education methods in Chinese traditional art education," Перспективы науки и образования, vol. 5, no. 71, pp. 713-728, 2024, doi: https://pnojournal.wordpress.com/2024-2/24-05/.

[19] F. Tao, "A new harmonisation of art and technology: Philosophic interpretations of artificial intelligence art," Critical Arts, vol. 36, no. 1-2, pp. 110-125, 2022, doi: 10.1080/02560046.2022.2112725.

[20] Y. Zhang, Y. He, Y. Xia, Y. Wang, X. Dong, and J. Yao, "Exploring the representation of Chinese cultural symbols dissemination in the era of large language models," International Communication of Chinese Culture, vol. 11, no. 2, pp. 215-237, 2024, doi: 10.1007/s40636-024-00293-z.

[21] M. C. Hung, R. Nakatsu, N. Tosa, and T. Kusumi, "Leartableluation based on psychological experiments," International Journal of Arts and Technology, vol. 14, no. 3, pp. 171-191, 2022, doi: 10.1504/IJART.2022.128444.