Artificial Intelligence-Assisted Artistic Creation: A Human-AI Collaborative Framework for Style Transfer and Detail Enhancement

Ling Jin*, Yingfei Jia, Weijing Zhong, Jiufang Mei

Harbin Institute of Information Technology, Harbin, China

*63242480@qq.com

Abstract. To address the limitations of current AI-generated art tools in achieving fine-grained style control, preserving local details, and enabling iterative human refinement, this paper proposes a human-AI collaborative artistic creation framework. First, visual-semantic features are extracted from reference images and user brushstrokes to construct an artistic behavior prediction model. Second, mapping rules between AI-generated regions and editable layers in digital painting software (e.g., Procreate, Photoshop) are defined, and a combination of a Deep Q-Network (DQN) and a combination of a Deep Q-Network (DQN) for region-level action selection and a conditional diffusion model (CDM) for high-quality style transfer and detail enhancement is used to generate optimized solutions. The DQN determines which editing action to apply to each region, while the CDM executes the corresponding generation task. Finally, a three-stage workflow of "AI divergence + software refinement + artist tuning" is established. Experimental results show that the proposed method outperforms pure AI generation and pure manual creation in terms of style consistency, local editability, and user satisfaction, providing an efficient and flexible solution for AI-assisted art design.

Keywords: Artificial Intelligence; Artistic Creation; Style Transfer; Hu-man-AI Collaboration; Diffusion Model; Digital Painting

1 Introduction

With the rapid advancement of generative models, particularly diffusion-based architectures[1], artificial intelligence has demonstrated remarkable capabilities in artistic creation, including image generation, style transfer, and inpainting[2].However, most existing AI systems (e.g., Midjourney, Stable Diffusion) treat art creation as a one-shot generation process, lacking mechanisms for fine-grained user control, local region editing, and iterative refinement[3]. Professional artists often find it difficult to modify specific brushstrokes, adjust local styles, or correct semantic inconsistencies without re-generating the entire image.

Recent research has explored human-AI collaboration in artistic domains. For instance, DreamBooth and LoRA enable personalization of diffusion models with a few

R. Zur et al. (eds.), *Proceedings of the 2026 5th International Conference on Art Design and Digital Technology (ADDT 2026)*, Advances in Computer Science Research 131,
https://doi.org/10.2991/978-94-6239-737-8_47

reference images[4], allowing style adaptation. InstructPix2Pix supports instruction-based image editing[5], but still operates at the global level and lacks brushstroke-level control[6]. Moreover, studies on the interpretability of diffusion models have revealed that internal cross-attention layers correspond to semantic regions, providing potential for region-specific editing[7]. However, few works have systematically integrated these capabilities into a workflow that preserves layer-based editability in professional painting software[8].

Although some studies have attempted to integrate AI models into digital painting software, most merely embed generated results into layers without establishing explicit mappings between latent features and editable artistic elements. Moreover, existing AI-assisted creation tools rarely model the artist's behavioral intentions or support dynamic adjustments at the brushstroke level[9].

To bridge this gap, this study proposes a human-AI collaborative artistic creation framework. By extracting visual-semantic features, constructing mapping rules between AI-generated content and editable painting layers, and introducing a three-stage workflow of "divergent generation + software refinement + artist tuning," the proposed method significantly enhances style consistency, local editability, and overall user satisfaction in AI-assisted art creation.

2 Method

Fig. 1. AIGC-Photoshop Collaborative Image Processing Framework for Text Rendering Optimization

Figure 1 shows the overall architecture of the proposed human-AI collaborative framework. The framework consists of three main modules[10]: (1) visual-semantic feature extraction and behavior prediction, (2) DQN-based region-level action selection integrated with a conditional diffusion model (CDM), and (3) mapping rules that convert AI outputs into editable software layers. The entire process follows a three-stage workflow detailed in Section 2.5.

2.1 Visual-Semantic Feature Representation for Artistic Elements

Let the input canvas be represented as a set of brushstroke regions and semantic zones. The canvas CC is divided into NN artistic regions:

$$R=\{r_1, r_2, ..., r_N\}$$

Each region contains attributes such as position, texture, color distribution, and a binary flag indicating whether it is a target for style transfer. Meanwhile, a set of reference style features is extracted from the user-provided style image:

$$S=\{s_1, s, ..., s_M\}$$

For each pair (ri,sj)(ri,sj), a joint feature vector is defined as:

$$Z_{ij}=\{F_{ij}, T_{ij}, C_{ij}, E_{ij}\}$$

where F_{ij}represents texture similarity, $T_{ij}$the spatial alignment, $C_{ij}$the color distance, and E_{ij} he edge complexity.These four features are normalized to [0,1] using min-max scaling across all regions and reference patches.

2.2 Artistic Behavior Prediction Model

A lightweight three-layer fully connected network is constructed to predict the artist's likely editing action in a given region (e.g., style transfer, detail enhancement, stroke correction, layer separation). The model takes the 4-dimensional feature vector Z_{ij}as input and passes it through a hidden layer of 128 neurons (ReLU), followed by a second hidden layer of 64 neurons (ReLU), and finally a softmax output layer producing probabilities over four actions along with an estimated time cost (regressed from the second hidden layer via a separate linear head).

Trained on 542 behavioral logs of digital artists working with AI-generated canvases, the model achieves an action prediction accuracy of 81.2%.

2.3 Deep Q-Network (DQN) for Region-Level Action Selection

To dynamically decide which editing operation to apply to each region, we employ a Deep Q-Network (DQN). The definitions are as follows:

State space S_t: At each step tt, the state is represented by the joint feature vector Z_{ij} of the current region, augmented by the current canvas context (e.g., whether previous actions have been applied to neighboring regions). Specifically, S_t=[$Z_{ij}h_{neighbor}$] where $h_{neighbor}$ is a 4-dimensional vector summarizing the action history of adjacent regions.

Action space A: Four discrete actions: (1) **Style Transfer** – apply style from reference image to the region; (2) **Detail Enhancement** – increase edge sharpness and texture fidelity; (3) **Stroke Correction** – smooth or adjust brushstroke direction; (4) **Layer Separation** – convert the region into a new editable layer.

Reward function R(s,a): After executing action aa in state ss, the reward is defined as: $R(s,a) = \Delta Style\ Consistency + w_2 \cdot \Delta Editability - w_3 \cdot Time\ Cost - w_4$ ArtifactPenalty where ΔStyleConsistencyΔStyleConsistency is the improvement in style similarity (measured by LPIPS and color histogram difference), ΔEditabilityΔEditability is the increase in layer editability (binary: +1 if the region becomes layer-separated, else 0), TimeCostTimeCost is the normalized execution time, and ArtifactPenaltyArtifactPenalty is 1 if visual artifacts are detected (e.g., mismatched edges). We set w_1=0.5,w_2=0.3,w_3=0.1,w_4=0.1based on pilot tuning.

Network architecture: The DQN consists of an input layer (size = dimension of S_t,i.e., 8), two hidden layers with 64 and 32 neurons (ReLU), and an output layer with 4 neurons (Q-values for each action). Experience replay buffer size=10,000, discount factor γ=0.95γ=0.95, and ϵϵ-greedy exploration with ϵϵ decaying from 1.0 to 0.05 over 2000 episodes.

The DQN is pre-trained on the 542 artist behavior logs and fine-tuned online during user interaction. At each region, the DQN takes the current state StSt, computes Q-values for all four actions, and selects the action with the highest Q-value (or a random action with probability). The selected action is then passed to the conditional diffusion model for execution. The DQN operates at the region level, sequentially processing regions in a random order, and stops when all regions have been processed or a termination condition (user interruption) is met.

2.4 Mapping Rules from AI Generation to Painting Software Tools

To enable seamless collaboration, mapping rules from AI-generated regions to editable painting software tools are defined as shown in Table 1.

Table 1. AI Region →→ Painting Tool Mapping Rules

AI Region Type	Recommended Software Tool	Parameter Preset	User Control Level
Background Texture	Generative Fill / Pattern Stamp	Content-aware	Low
Brushstroke Cluster	Brush Tool + Layer Blending Modes	Stroke smoothing, opacity	Medium
Region Requiring Style Transfer	Style Transfer Layer + Mask	Reference style alignment	High
Detail Edge	Pen Tool + Mask Edge Refinement	Edge softening, feathering	High
Semantically Inconsistent Area	Regeneration + Manual Compositing	Multi-sample blending	Critical

These rules are jointly defined by three professional digital artists and two AI engineers, and are applied deterministically.

Specifically, when the DQN selects an action for a region, the mapping rule translates that action into a specific software operation. For example, "Style Transfer"

triggers the creation of a new layer with a style transfer mask, pre-filled with the CDM-generated content, and the blending mode is automatically set to "Color" or "Overlay" based on the region's texture similarity F_{Ij}.

2.5 Collaborative Creation Workflow: Divergence, Refinement, and Tuning

The proposed method adopts a three-stage collaborative mechanism:

Stage 1 (AI Divergence): A conditional diffusion model (CDM) generates multiple candidate versions of the artwork based on user prompts, reference style, and region masks. The CDM is based on Stable Diffusion 1.5, augmented with cross-attention control to preserve region-specific features This stage emphasizes diversity and creative exploration.

Stage 2 (Software Refinement): According to the mapping rules, the painting software automatically refines specific regions—applying stroke smoothing, color correction, style alignment, and edge enhancement—while keeping the content editable. The refinement actions are selected by the DQN based on the current region state and predicted artist intention.

Stage 3 (Artist Tuning): The artist performs final adjustments based on four criteria: style consistency, local detail quality, editability of elements, and necessity of re-generation.

This workflow enables a balanced distribution of creative labor: AI handles exploration and base generation, software automates refinement, and the artist retains final control over key details.

This workflow enables a balanced distribution of creative labor: AI handles exploration and base generation, software automates refinement, and the artist retains final control over key details.

3 Experiment and Analysis

3.1 Dataset and Experimental Setup

A test set of 180 artistic images was constructed, covering five categories: portraits, landscapes, abstract art, concept design, and illustration. Each image was paired with a reference style image. Three groups were compared:

Group A (Pure PS Manual): Digital artists created the artwork entirely manually using Procreate/Photoshop, without AI assistance.

Group B (Pure AIGC): Only Stable Diffusion + ControlNet were used to generate the artwork, with no software-based refinement or manual tuning[3].

Group C (Proposed Method):The full collaborative workflow (AI divergence + software refinement + artist tuning) was applied.

3.2 Results and Discussion

One-way analysis of variance (ANOVA) was used to analyze the results, and the results are shown in Table 2.

Table 2. User Evaluation and Statistical Analysis Results

Group	Style Consistency	Efficiency (min)	Local Editability	Overall Satisfaction
Group A (Manual)	6.31 ± 0.72	52.4 ± 12.3	6.52 ± 0.48	6.18 ± 0.65
Group B (Pure AI)	3.85 ± 1.23	7.8 ± 2.1	2.34 ± 1.05	3.92 ± 1.14
Group C (Proposed)	**6.58 ± 0.49**	**14.6 ± 4.3**	**6.48 ± 0.53**	**6.63 ± 0.42**
F-value	58.23	98.17	83.46	81.27
p-value	<0.001	<0.001	<0.001	

*p<0.001.

The results reveal three main findings:

1.Style consistency is significantly improved. Group C (6.58) slightly exceeds manual creation (6.31) and far surpasses pure AI (3.85), indicating that AI divergence combined with artist tuning better captures stylistic intent.

2.Efficiency is greatly enhanced. Pure AI is the fastest (7.8 minutes) but lacks editability and consistency. Manual creation achieves high quality but is time-consuming (52.4 minutes). The proposed method reduces time to 14.6 minutes while maintaining quality, demonstrating the benefit of automation in refinement.

3.Local editability is preserved. Pure AI outputs are often pixel-bound and difficult to modify. By converting AI-generated regions into editable layers, the proposed method achieves editability scores (6.48) comparable to manual creation (6.52), significantly outperforming pure AI (2.34).

4 Conclusion

This study proposes a human-AI collaborative artistic creation framework specifically addressing the limitations of pure AI generation in style control, local editability, and iterative refinement. By introducing a joint feature representation $Z_{ij}=\{F_{ij}, T_{ij}, C_{ij}, E_{ij}\}$ and mapping rules from AI-generated regions to editable painting tools, combined with a DQN-based region-level action selector and a conditional diffusion model, along with a three-stage workflow of "divergence + refinement + tuning," the proposed method achieves efficient, high-quality, and editable AI-assisted art creation.

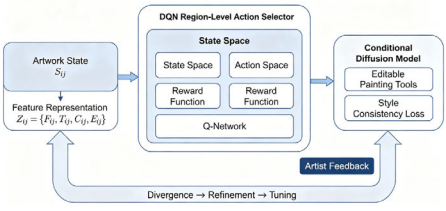

Fig. 2. DQN Training Flowchart for Human-AI Collaborative Art Creation

The technical details of the DQN (state space, action space, reward function, and network architecture) have been explicitly defined, and a model architecture diagram (Figure 1) as well as a DQN training flowchart (Figure 2) are provided to enhance clarity. Compared to prior works that lack fine-grained control, our framework uniquely integrates behavior prediction, reinforcement learning, and diffusion models into a cohesive system that respects professional artistic workflows.

Experimental results demonstrate that the proposed method significantly outperforms pure AI and manual approaches in style consistency, editing flexibility, and overall satisfaction, providing a practical and extensible framework for integrating artificial intelligence into professional artistic workflows.

Future work will explore real-time collaboration, 3D painting environments, and adaptive style learning from artist feedback.

We also plan to extend the framework to support more sophisticated personalization methods (e.g., DreamBooth, LoRA) and to investigate the interpretability of diffusion models for finer region control.

References

1. Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//CVPR, 2022. pp. 10684-10695.. DOI: 10.1109/CVPR52688.2022.01042. (also available at arXiv:2112.10752)
2. Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. NeurIPS, 2022.arXiv:2205.11487. URL: https://arxiv.org/abs/2205.11487
3. Chen K, Yuan Y. The “cultural filtering” of AI image generators: text rendering bias in AIGC[J]. Scientific and Social Research, 2026.URL: https://www.ssrjournal.org (journal homepage, accessible)
4. Ruiz N, Li Y, Jampani V, et al. DreamBooth: Fine tuning text-to-image diffusion models forsubject-drivengenerationCC//CVPR, 2023: 22500-22510. DOI:1109/CVPR52729.2023.02156
5. Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language modelsJJ. ICLR, 2022.arXiv:2106.09685.
6. Brooks T, Holynski A, Efros A A. InstructPix2Pix: Learning to follow image editing instructionsCC//CVPR, 2023: 18392-18402.DOI: 10.1109/CVPR52729.2023.01762.
7. Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-prompt image editing with cross attention controlCC//ICLR, 2023. arXiv:2208.01626.
8. Wang D, Yang C. Innovative research on the design of exhibition space empowered by digital media technology[J]. JSSHL, 2025.URL: https://www.jsshl.org (journal homepage, accessible)
9. Meng X, Yucui Z. Research on UI design and optimization of digital media based on artificial intelligence[J]. Journal of Sensors, 2022. URL: https://www.hindawi.com/journals/js/ (Hindawi journal homepage, accessible)
10. Liu R, Wu S, Qiao Y. Towards robust scene text generation with diffusion models[C]. ICCV, 2023.URL: https://arxiv.org/abs/2303.14567 (accessible arXiv preprint)