Research on Stylized Image Generation of Local Chronicles Ancient Books Based on LoRA Fine-tuning

Huabiao Li[1,2], Tianhang Liu[3], Bin Wang[3]

[1]School of Engineers, Zhejiang University, Hangzhou 310000, Zhejiang Province
[2]Data Management and Analysis Center, National Museum of China, Beijing 100000
[3]School of Electronic Information (School of Artificial Intelligence), Northwest University, Xi'an 710000, Shaanxi Province

Abstract: ***To improve the visualization and dissemination efficiency of the content of local chronicles ancient books, this paper explores an artificial intelligence painting generation scheme based on the Low-Rank Adaptation (LoRA) fine-tuning technology, aiming to assist the public to better understand and appreciate the ancient book content by generating intuitive and easy-to-understand illustrations. First, a selected dataset containing a small number of paintings in the style of local chronicles ancient books is constructed and refined annotation is carried out. Then, a pre-trained text-image generation model is used as the basic framework, and the LoRA technology is used to carry out targeted fine-tuning to improve the model's understanding ability of specific domain texts and the quality of image generation. Finally, the trained LoRA model is combined with the basic model for image generation inference. The experimental results show that this method can reproduce a specific ancient book painting style with high fidelity and accurately translate the text semantics, providing an efficient and low-cost technical path for the digital interpretation and dissemination of cultural heritage.***

1. INTRODUCTION

With the rapid development of artificial intelligence technology, its applications have penetrated into all aspects of social life, including multiple fields such as education, healthcare, and transportation. Especially in the field of humanities, the rise of large models has brought unprecedented opportunities for the protection and inheritance of cultural heritage. As an important branch of AIGC (Artificial Intelligence Generated Content), the Text-to-Image (T2I) generation model has shown great potential in the popularization and inheritance of cultural heritage such as local chronicles and ancient books due to its powerful creativity and expressiveness. The T2I model can automatically generate corresponding images according to the input text description, which provides a novel and intuitive way for the interpretation of local chronicles and ancient books. Specifically, by transforming the complex and difficult-to-understand text in ancient books into an easy-to-understand visual form, it can not only help ordinary readers obtain the historical information and cultural connotations in them more easily, but also arouse more people's interest in local chronicles and ancient books, thus promoting the wide spread of traditional culture.

Although general T2I models with large training sets perform well in various tasks, they have obvious limitations in generating images of specific styles. This is mainly because the training data of specific artistic styles is usually scarce and difficult to support the complete retraining of the model. Therefore, how to perform parameter-efficient fine-tuning (PETF) of general T2I models using a small number of samples has become a hot topic in current research.

Currently, a commonly used PEFT technique is the Low-Rank Adaptation (LoRA) fine-tuning method. In the generation of images in the style of local chronicles and ancient books, by collecting a small number of representative ancient book illustrations or related images and using the LoRA method to fine-tune the pre-trained T2I model, not only can the transfer of specific styles be achieved at a low computational cost, but also the generation quality and generalization ability of the model in the new style can be effectively improved. This study not only provides a new path for the digital dissemination of local chronicles and ancient books, but also explores feasible technical means for the protection and inheritance of other cultural heritages.

2. RELATED WORK

In the domain of structured data processing, Deng [1] enhanced neural network performance on tabular data through knowledge distillation and RankGauss transformation. Building on multimodal learning, Zi and Deng [2]

developed a joint modeling framework that integrates medical images and clinical text for early diabetes risk detection, highlighting the potential of AI in healthcare. In the automotive industry, Ziren [3] conducted a dynamic optimization and multi-regional performance validation of automotive sales strategies in the United States, providing data-driven insights for market-specific decision-making. The field of computer vision has also seen significant progress, as Peng et al. [4] proposed a lifelong domain adaptive approach for 3D human pose estimation, enabling robust performance across varying domains, while Zheng et al. [5] introduced Diffmesh, a motion-aware diffusion framework for human mesh recovery from videos, advancing video-based 3D reconstruction. In photonics, Tang et al. [6] investigated the design and optimization of shallow-angle grating couplers for vertical emission from Indium Phosphide devices, contributing to integrated optical component development. Sun [7] explored AI-assisted UI design, demonstrating how generative tools can enhance both efficiency and creativity in the design process. Within the financial sector, Yang et al. [8] designed a full-cycle intelligent risk control system for pre-loan, mid-loan, and post-loan lending, employing AI-driven closed-loop management to strengthen online credit security. Shen et al. [9] applied the Whale Optimization Algorithm to financial payment fraud detection, showcasing the utility of bio-inspired algorithms in anomaly identification. Extending this line of financial security research, Yang and Zhang [10] proposed an edge-enabled real-time fraud detection framework for network lending terminals operating under low-latency constraints, addressing the demand for instantaneous threat response. In digital marketing, Zhou [11] developed a digital precision distribution strategy for social media content on private domain platforms in the automotive industry, utilizing a collaborative filtering model based on user behavior to enhance content targeting. Wensi [12] examined AI-enabled data visualization marketing for automated production lines, focusing on building customer trust and improving lead-to-order conversion rates. In the realm of data management, Yang et al. [13] presented HGMatch, a match-by-hyperedge approach for subgraph matching on hypergraphs, offering a novel solution for complex graph analytical tasks. Ukey et al. [14] proposed an efficient method for continuous kNN join over dynamic high-dimensional data, tackling scalability issues in real-time data processing environments. Finally, Lian and Chen [15] investigated complex data mining analysis and pattern recognition based on deep learning, advancing foundational techniques for knowledge discovery.

3. METHOD

This paper adopts the widely concerned diffusion model technology at present, uses the pre-trained T2I model, and fine-tunes it through the LoRA method, aiming to generate images with specific styles. This method can not only effectively improve the quality of image generation, but also introduce new style features while maintaining the performance of the original model, which plays a promoting role in the popularization of local chronicles and ancient books.

3.1 Methods for Text-to-Image Generation

At present, most multimodal generation models tend to adopt diffusion models as the core architecture, mainly due to their ease of training and excellent generation quality. Based on the basic concept of diffusion models, the Denoising Diffusion Probabilistic Model (DDPM) was introduced in 2020, which is an important milestone in this field. In the forward diffusion process, DDPM gradually adds noise to the data, ultimately resulting in a Gaussian noise distribution. The transformation of data from x_0 to x_1 at each time step t can be described by Equation (1).

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon_t. \tag{1}$$

In the denoising diffusion process, the model learns to recover the original data from the noise. The generation process is described by Equation (2):

$$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta (x_t, t)) \tag{2}$$

where μ_θ and $\sum_\theta$ are the mean and variance parameters learned by the model. The diffusion process gradually adds Gaussian noise to the original image in a fixed Markov chain, ultimately converting the image into Gaussian noise. The inverse process then gradually recovers the original image through denoising to achieve image generation. This method is renowned for its excellent generation performance. It provides new insights for the T2I task by using the diffusion process to gradually convert the text description into a high-quality image.

Stable Diffusion is the first model to apply the diffusion model in the latent space of a powerful pre-trained autoencoder. It has gained wide popularity due to its open-source nature and excellent community ecosystem. It follows a two-stage approach: in the first stage, the image is compressed into a latent representation to reduce computational complexity; in the second stage, the DDPM structure is used. Additionally, a cross-attention

mechanism is introduced, allowing the text or image sketch to influence the diffusion model and generate the desired image.

Stable Diffusion consists of three main components: the Variational Autoencoder (VAE), the text encoder of Contrastive Language-Image Pretraining (CLIP), and the U-Net noise prediction network. The VAE is responsible for mapping the image to the latent space, which not only reduces the required memory consumption but also improves computational efficiency, making it possible to generate and train large images under existing hardware conditions. The U-Net consists of multiple ResNet blocks, Transformer blocks, upsampling and downsampling layers, and learns how to reconstruct the image from the noise by training on text-image pairs. The text encoder of CLIP is used to encode the input text, which is passed to the U-Net as a condition to guide the image generation process through the cross-attention mechanism. Finally, the U-Net takes random noise and the given text to obtain a latent variable, and finally uses the VAE as the bridge between the latent space and the image to generate an image that highly matches the text description. This model has undergone multiple iterations, such as Stable Diffusion v1.5, v2.1, and XL versions. In this study, Stable Diffusion v1.5 was selected as the base model.

3.2 LoRA Fine-tuning Method

LoRA is essentially an approximate numerical factorization technique for low-rank factorization of the feature matrix. Through this technique, the number of parameters of the feature matrix can be significantly reduced. Its core idea is that during the fine-tuning process, the weight update matrix ΔW of the pre-trained model has a low "intrinsic rank", so it can be approximated by the product BA of two low-rank matrices, that is, $\Delta W \approx BA$. Specifically, during training, first freeze the weights of the Stable Diffusion model to ensure that most of the parameters of the pre-trained model remain unchanged. Then, in the U-Net network of the Stable Diffusion model, for the weight matrix of the Cross-Attention module, connect a trainable branch composed of low-rank matrices A and B in parallel. Finally, only the parameters of the Cross-Attention part are fine-tuned.

In this study, Stable diffusion v1.5 is used as the pre-trained model. To keep the original model weights unchanged when introducing new training data, it is necessary to freeze Stable Diffusion, and then insert trainable weights beside the Text-Encoder of CLIP and the cross-attention layer of U-Net. Finally, only matrices A and B are trained. Figure 1 shows the LoRA method used in this paper.

4. EXPERIMENTAL METHODS

In this study, we adopted the LoRA fine-tuning strategy, using the long-standing local chronicles literature "The Gazetteer of Shuntian Prefecture" as the content basis, aiming to generate images with the style of traditional ancient books. To construct prompts suitable for the T2I model, we first extracted descriptive text fragments from the "Geography Section of The Gazetteer of Shuntian Prefecture". Subsequently, we used the ChatGLM large language model to modernize and artistically expand these ancient text fragments, generating a series of "popular science narrative texts" that not only retain the original information but also conform to modern semantic habits as input prompts for image generation. A series of highly relevant popular science narrative texts were formed, and these texts were used as experimental media to drive the image generation process.

4.1 Training Data

The dataset used for training in this experiment is the Chinese-Landscape-Painting-Dataset, which contains 2,192 high-quality Chinese landscape paintings. After screening, we selected 246 images from it as the training set. We first labeled the training set images and set specific stylization prompts [U] in each image, such as the Shun-TianFuZhi Style, in order to activate the learned specific style during training and inference.

4.2 Training Details

The pre-trained model used during training is Stable Diffusion v1.5, and the training parameters are as follows: the training resolution is 512×512, max_train_epochs is 15, batch_size is 2, learning_rate is. To save memory and reduce computational costs, the optimizer selects AdamW 8-bit, and the number of training steps is 12120.

4.3 Inference Details

We adopted the "Geography" section in "The Gazetteer of Shuntian Prefecture" and extracted a part of the text

content. After translation, understanding, and generation by the ChatGLM model, we obtained prompts that can be used to generate images.

During inference, we used Stable Diffusion v1.5 as the base model, inserted the trained LoRA model, set the sampler to DDIM, the sampling steps to 35, the CFG (Classifier Free Guidance) to 7, and the input text to "[U], prompt", where "prompt" is

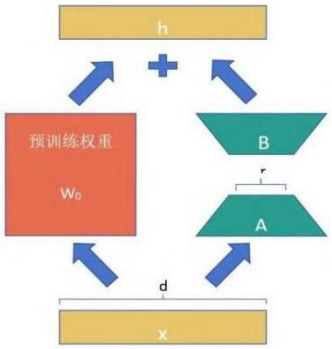


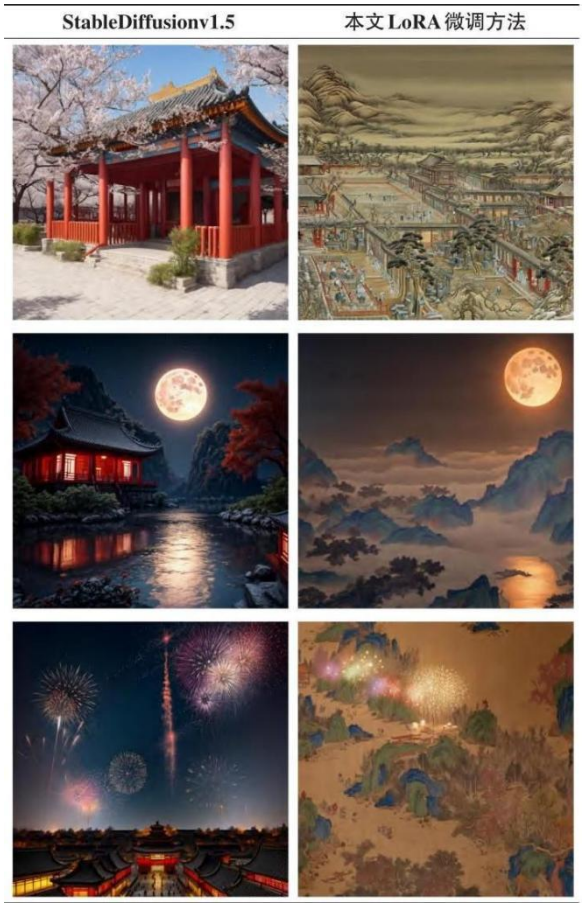
Figure 1: Schematic Diagram of the LoRA Method

Popular science narrative text generated based on "The Gazetteer of Shuntian Prefecture".

5. EXPERIMENTAL RESULTS

We respectively used the open-source Stable Diffusion v1.5 model, as well as LoRA trained with the training set combined with the base model to generate several images for each prompt under the same parameters, as shown in Table 1. The mapping scheme generally outperforms existing methods. Future work will focus on co-designing this framework with the initial layout algorithm, and will mainly study generating an initial mapping that can "perceive the future" based on the subgraph isomorphism algorithm or reinforcement learning method, so that the search process can start from a more favorable position, and thus有望 further compress the total number of SWAP gates required for the mapping and approach the theoretical optimal solution of the problem without significantly increasing the search overhead.

Table 1: Generate several images



The research conducts a systematic multi-dimensional evaluation of the generated images. For dimensions such as the style consistency, the graphic-text matching degree, the detail richness, and the artistic beauty of the images, 4 art practitioners are asked to score. The results show that this method has the highest score, as shown in Table 2.

Table 2: Multi-dimensional Evaluation Score Table of Images

Evaluation dimension	Stable Diffusion v1.5 (Average score + Standard deviation)	LoRA fine-tuning method (Average score + Standard deviation)
Style consistency	2.1±0.6	4.3±0.4
Image-text matching degree	2.3±0.5	4.5±0.3
Detail richness	2.0±0.7	4.2±0.5
Artistic beauty	2.2±0.6	4.4±0.4
Comprehensive score (Mean of each dimension)	2.15±0.55	4.35±0.38

It can be seen from the results of the comparative experiment that for the same text description, when only using the open-source model, the generated images have a large gap from the task objectives in terms of both style and details. Therefore, it is necessary to fine-tune the pre-trained diffusion model. The images generated by the LoRA method after training are closer to the training set and the task objectives in terms of style and effect. The mountain and water levels are distinct, the画面 is more consistent with the text semantics, the details such as brushstrokes have good effects, and the overall layout is similar to real works, winning the affirmation of more art practitioners.

6. CONCLUSION

This paper innovatively applies the LoRA fine-tuning technique to the task of generating illustrations in the style of local chronicles and ancient books, explores its application potential in this field, and achieves efficient and targeted model optimization by fine-tuning pre-trained diffusion models, significantly improving the quality and diversity of the generated illustrations. This method fully utilizes the efficient training and high-quality generation capabilities of diffusion models, successfully generates illustration works that are both scientifically accurate and artistically beautiful, and significantly enhances readers' understanding and interest in the content of local chronicles and ancient books. Future research can further expand the diversity of the dataset to cover the styles of

local chronicles in different periods and regions, and explore the possibility of combining this technology with interactive digital museums and educational applications.

REFERENCES

[1] Deng, Xiaoxiao. "Enhancing Neural Network Performance on Tabular Data via Knowledge Distillation and RankGauss Transformation." 2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2025.

[2] Zi, Yun, and Xiaoxiao Deng. "Joint modeling of medical images and clinical text for early diabetes risk detection." Journal of Computer Technology and Software 4.7 (2025).

[3] Ziren, Z. (2026). Dynamic Optimization and Multi-Regional Performance Validation of Automotive Sales Strategies in the United States. Academic Journal of Natural Science, 3(1), 1-7.

[4] Peng, Qucheng, Hongfei Xue, Pu Wang, and Chen Chen. "Lifelong Domain Adaptive 3D Human Pose Estimation." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 10, pp. 8358-8366. 2026.

[5] Zheng, Ce, et al. "Diffmesh: A motion-aware diffusion framework for human mesh recovery from videos." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

[6] Tang, Yingheng, et al. "Design and Optimization of Shallow-Angle Grating Coupler for Vertical Emission from Indium Phosphide Devices." (2020).

[7] Sun, Lingxin. "AI-Assisted UI Design: Enhancing Efficiency and Creativity through Generative Tools." Journal of Computer Technology and Applied Mathematics 3.1 (2026): 19-27.

[8] Yang, X., Xue, H., Hu, Q., & Zhang, Y. (2025, October). Design of a full-cycle intelligent risk control system for pre-loan, mid-loan, and post-loan lending: AI-driven closed-loop management of online credit security. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 1022-1027).

[9] Shen, Zepeng, et al. "Research on Application of Whale Optimization Algorithm in Financial Payment Fraud Detection." 2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID). IEEE, 2025.

[10] Yang, X., & Zhang, Y. (2026). Edge-Enabled Real-Time Fraud Detection for Network Lending Terminals under Low-Latency Constraints. Journal of Computer Technology and Applied Mathematics, 3(1), 55-62.

[11] Zhou, Z. (2025, November). Digital precision distribution strategy for social media content on private domain platforms in the automotive industry: a collaborative filtering model based on user behavior. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 516-521).

[12] Wensi, L. (2026). AI-Enabled Data Visualization Marketing for Automated Production Lines: Building Customer Trust and Improving Lead-to-Order Conversion. Academic Journal of Natural Science, 3(1), 8-13.

[13] Yang, Z., Zhang, W., Lin, X., Zhang, Y., & Li, S. (2023, April). HGMatch: A Match-by-Hyperedge Approach for Subgraph Matching on Hypergraphs. In 2023 IEEE 39th International Conference on Data Engineering (ICDE) (pp. 2063-2076). IEEE.

[14] Ukey, N., Zhang, G., Yang, Z., Li, B., Li, W., & Zhang, W. (2023). Efficient continuous kNN join over dynamic high-dimensional data. World Wide Web, 26(6), 3759-3794.

[15] Lian, J., & Chen, T. (2024). Research on Complex Data Mining Analysis and Pattern Recognition Based on Deep Learning. Journal of Computing and Electronic Information Management, 12(3), 37-41.