Full-Page Chinese Calligraphy Generation via LoRA Fine-Tuning of Stable Diffusion

Zhiyin Pan
zhiyinp@stanford.edu

Huici Pan
huici@stanford.edu

Jieshu Huang
jieshuh@stanford.edu

Abstract

We present a novel approach for generating full-page Chinese calligraphy using Low-Rank Adaptation (LoRA) to fine-tune Stable Diffusion, a large-scale latent diffusion model. Unlike prior work focused on individual character synthesis or supervised style transfer, our method enables stylistically coherent, layout-aware image generation without character-level labels. By applying LoRA to both attention and convolution layers, we achieve efficient domain adaptation using a small dataset of historically diverse calligraphy samples. Our experiments show that fine-tuning enables the model to reproduce brush textures, spatial rhythm, and stylistic traits aligned with specific calligraphers. In addition, we discovered that the model struggles with semantic accuracy and nuanced script differentiation due to limited data and CLIP' s short input length. Despite these constraints, our method demonstrates the potential for adapting diffusion models to low-resource, culturally significant visual domains with minimal supervision.

1. Introduction

Diffusion models have emerged as a powerful generative framework, achieving state-of-the-art results in image synthesis, inpainting, and conditional generation. Among these, denoising diffusion probability models (DDPM) by Ho *et al*. [3] form the foundational architecture, where image generation is modeled as a gradual denoising process from Gaussian noise. Stable diffusion introduced by Rombach *et al*. [13], a latent diffusion model built on DDPM, has demonstrated both high quality generation and computational efficiency.

Although these models are powerful, fine-tuning them for domain-specific image generation tasks remains a challenge because of their large size and training cost. Low-Rank Adaptation (LoRA) by Hu *et al*. [4] inserts trainable rank decomposition matrices into existing model weights, allowing new capabilities to be learned with minimal changes to the original model parameters. This makes LoRA particularly well-suited for adapting generative models like Stable Diffusion to niche or underrepresented domains.

Chinese calligraphy is a visually rich art form marked by expressive brushwork and complex composition. Although recent studies such as *calliffusion* by Liao *et al*. [6] and *Moyun* by Liu *et al*. [8] explore the diffusion-based generation of individual characters or style transfer, they are based on character-level control or recognition-based models.

In contrast, our goal is to generate full-page calligraphy compositions without explicit supervision, allowing the model to freely learn both stylistic and structural patterns from example images. Importantly, our method does not train a separate model for each calligraphy style. Instead, we fine-tune a single base model using LoRA with a dataset containing multiple genres of calligraphy, each associated with a descriptive prompt (e.g., ”mi fu xingshu calligraphy” or ”chu suiliang kaishu calligraphy”). During inference, the desired genre or style is specified as part of the text prompt, which acts as a control signal to guide generation. This prompt-conditioned generation enables the model to synthesize images in various calligraphic styles using the same underlying model weights. Thus, our method offers a scalable and flexible framework for modeling diverse artistic styles without requiring separate fine-tuning for each one.

We show that, with minimal domain-specific supervision, Stable Diffusion can be guided to generate high-fidelity, stylistically coherent calligraphy compositions across different scripts. Our findings contribute to both the technical development of efficient generative fine-tuning methods and the cultural enrichment of generative AI applications.

2. Related Work

2.1. Stable Diffusion

Stable Diffusion [13] is a latent text-to-image generative model that has become a foundational method for controllable image synthesis. It comprises three key components: (1) a variational autoencoder (VAE) that maps images to a compressed latent space $\mathcal{Z}$, (2) a UNet-based denoising model that predicts noise residuals during diffusion steps in

$\mathcal{Z}$, and (3) a text encoder derived from CLIP [11], which maps text prompts into conditioning vectors. Generation is guided by the objective:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\epsilon,\mathbf{z}_0,t}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}))\right\|_2^2\right], \tag{1}$$

where ϵ is the added noise, $\mathbf{z}_t$ is the latent at timestep t, and $\mathbf{c}$ is the conditioning text embedding.

Because Stable Diffusion operates in a lower-dimensional latent space instead of pixel space, it achieves higher efficiency and scalability. These properties, along with its open-source availability, make it well-suited for lightweight adaptation methods such as Low-Rank Adaptation (LoRA). Our work builds on Stable Diffusion to study the generation of full-page Chinese calligraphy compositions, leveraging prompt-guided fine-tuning to control genre and style through text prompts.

In the text-to-image setting, conditioning is introduced through cross-attention, where the noise prediction function becomes $\epsilon_\theta(\mathbf{z}_t, t, \tau(\mathbf{c}))$, with $\tau(\mathbf{c})$ denoting a text embedding. Our project builds on this framework by adapting cross-attention layers using task-specific LoRA.

2.2. Parameter-Efficient Fine-Tuning with LoRA

To efficiently adapt large-scale diffusion models, we leverage Low-Rank Adaptation (LoRA) [4], which reparameterizes weight updates in a low-rank form:

$$W' = W + \alpha AB, \tag{2}$$

where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ are learned matrices with rank $r \ll \min(d, k)$, and α is a scaling factor. This allows for efficient fine-tuning of specific submodules (e.g., attention or convolution layers) while keeping the large base model frozen.

Recent advances in LoRA extend the original method in several directions, such as adaptive rank scaling [16], LoraHub [5], and integration with parameter-efficient adapters [10]. These works target general-purpose language or vision-language models like LLMs or CLIP [11].

However, for diffusion-based image generation tasks, the original LoRA remains a widely used and strong baseline. Studies such as `LoRA-Diffusion` [9] and `DreamBooth-Lora` [15] demonstrate that even vanilla LoRA, when properly targeted to specific submodules (like attention layers in U-Net), can enable fast and high-quality personalization or style transfer with limited data.

In our work, we adopt this vanilla LoRA framework and extend its application to not only the linear layers of attention mechanisms but also to selected convolution layers within the U-Net of Stable Diffusion. We find this approach sufficient for adapting the model to complex, high-resolution visual domains such as Chinese calligraphy. Given the limited availability of training data (around 350 samples), the simplicity and robustness of the original LoRA make it especially appealing for our setting.

2.3. Diffusion for Chinese Calligraphy

Some prior work has explored the use of diffusion models for Chinese calligraphy, particularly at the character level. *CalliFusion* [6] proposes a diffusion-based model that generates individual Chinese characters in specific calligraphy styles. Their method relies on glyph-level supervision and assumes access to structurally aligned character datasets, making it suitable for applications such as font generation or calligraphy style transfer. Similarly, *MoYun* [8] focuses on style-specific calligraphy generation by learning to transfer style attributes from a reference character image to a standard printed glyph. These approaches are effective for producing high-quality stylized characters but are limited in scope to single-character synthesis and require fine-grained alignment between visual style and character identity.

In contrast, our method targets full-page calligraphy composition without explicit character-level supervision. Rather than generating isolated characters, we aim to synthesize entire compositions that reflect the aesthetic structure, brush dynamics, and layout characteristic of historical calligraphy works. Our approach leverages LoRA-based fine-tuning on top of a pre-trained text-to-image diffusion model, using weakly paired image-text data where the textual prompt provides only high-level style cues (e.g., artist and script type). This setup enables more flexible adaptation across styles and supports open-ended generation tasks not constrained by glyph-level supervision or paired data.

3. Data

We curated a custom dataset focused on historical and stylistically diverse Chinese calligraphy, comprising high-resolution images of complete works with associated style metadata. This was essential, as no existing public dataset meets the needs of our task.

Most available datasets contain isolated character images, useful for recognition or stroke analysis but inadequate for capturing the full-page composition and stylistic flow of an artist’s work. Full-page calligraphy in digital form is rare and often lacks consistent labeling or metadata, making it unsuitable for training generative models on coherent, page-level styles.

We collected 139 full-page samples of Mi Fu’s (米芾) xingshu and 210 samples of Chu SuiLiang (褚遂良）kaishu styles from reputable online archives and museums. Each image was paired with a manually curated .txt file including the artist’s name, style, and a brief description. This ensured strong semantic alignment between text and image—crucial for training text-to-image models like Stable Diffusion.

To enhance stylistic diversity, we added 63 samples from Wen Zhengming (文征明) and 24 from Zhao Mengfu (赵孟頫), both representing distinct traditions. All images were annotated with .txt metadata for use as conditioning prompts during training and inference.

We applied standard preprocessing, including resizing images to 1024×1024 for compatibility with Stable Diffusion. As the pretrained model struggled with Chinese tokens, we translated style and author labels into English for training data and test prompts, enabling generation using the model’s existing vocabulary.

	English prompt	Chinese prompt
Mi Fu’s (米芾) 行书	139	139
Chu Suiliang’s (褚遂良) 楷书	210	210
Zhao Mengfu (赵孟頫) 楷书	24	24
Zhengming (文征明) 楷书	63	63
Total	**348**	**348**

Table 1: Training corpus size overview

4. Methods

4.1. Approach

Our work treats calligraphy generation as a holistic image synthesis task. Full-page calligraphy places emphasis on the flow, rhythm, spatial composition, and emotional expressiveness of the artwork, qualities that go beyond individual characters. We deliberately avoid recognition constraints and allow the model to generate expressive, layout-aware imagery with stylistic flexibility.

To the best of our knowledge, our work is the first to explore full-page Chinese calligraphy generation using LoRA-fine-tuned latent diffusion models without explicit character-level supervision.

Diffusion Models and Lightweight Fine-Tuning Our approach leverages Stable Diffusion by Rombach [13], a latent DDPM model, and applies Low-Rank Adaptation (LoRA) Hu *et al*. [4] for efficient fine-tuning. LoRA enables domain adaptation with minimal memory and compute cost by injecting trainable low-rank updates into the attention layers.

4.2. Baseline Method

The baseline in our study is the original Stable Diffusion model from the Hugging Face library, as introduced by Rombach *et al*. [14], without any fine-tuning. This serves to illustrate the limitations and generic nature of image generation when no domain-specific adaptation is applied.

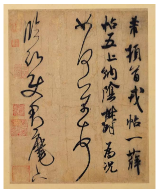
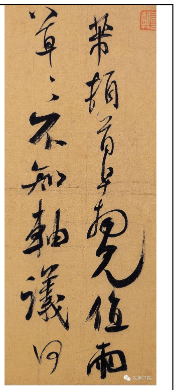
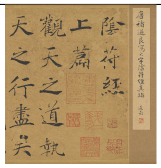
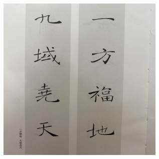
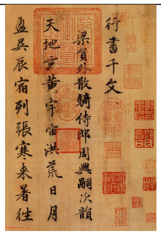
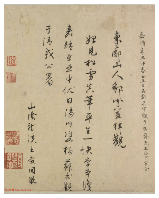
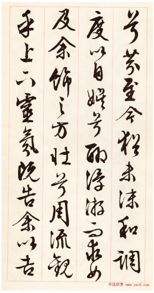
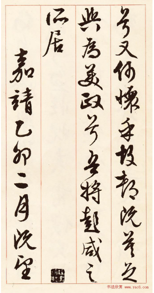
Mi Fu 米芾行书		
Chu Suiliang 褚遂良楷书		
Zhao Mengfu 赵孟頫楷书		
Zhengming 文征明楷书		

Table 2: Sample training image from each calligrapher’s style.

4.3. LoRA Injection to Stable Diffusion Model

U-Net Architecture and LoRA Injection The core denoising network in Stable Diffusion is a **U-Net**, which processes latent image representations at multiple resolutions through a sequence of *downsampling*, *bottleneck*, and *upsampling* stages. Each stage includes a combination of:

- **Residual blocks** with 2D convolutional layers
- **Self-attention** blocks (at selected resolutions)
- **Cross-attention** blocks (for text conditioning)

This U-Net is the main target for LoRA-based fine-tuning, as it contains the bulk of the learnable parameters responsible for image synthesis.

Attention-Based LoRA Injection We first apply LoRA to the *cross-attention layers*, which follow the standard attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

with linear projections defined as:

$$Q = XW^Q, \quad K = YW^K, \quad V = YW^V \tag{4}$$

LoRA is applied to the query and value projections:

$$W^Q_{\text{adapted}} = W^Q + \Delta W^Q = W^Q + A_Q B_Q, \tag{5}$$

$$W^V_{\text{adapted}} = W^V + A_V B_V \tag{6}$$

This allows the model to adjust its text-to-image alignment and spatial attention without modifying the base model weights.

Convolutional LoRA Injection Beyond attention, we also experiment with LoRA applied to *2D convolutional layers* in the residual blocks. Given a convolutional layer with weights $W_{\text{conv}} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k \times k}$, we reshape it into a matrix and inject a LoRA update:

$$W'_{\text{conv}} = W_{\text{conv}} + A_{\text{conv}} B_{\text{conv}} \tag{7}$$

where:

$$A_{\text{conv}} \in \mathbb{R}^{C_{\text{out}} \times r}, \\ B_{\text{conv}} \in \mathbb{R}^{r \times (C_{\text{in}} \cdot k^2)}$$

The result is reshaped back to the original 4D convolution weight shape. This enables LoRA to influence *low-level visual features* such as brush stroke texture, edge dynamics, and local contrast —all essential for expressive Chinese calligraphy.

Configuration Comparison We evaluated several configurations:

- **Attention-only LoRA:** Injected into query/value of cross-attention layers
- **Conv-only LoRA:** Injected into selected convolutional layers
- **Combined LoRA:** Applied to both

4.4. Inference

At inference time, we generate full-page images using prompts specifying style or author, e.g., "Chinese calligraphy, style: xingshu, artist: Mi Fu". Importantly, we do **not** require character-level supervision or recognition, allowing the model to generate aesthetically coherent compositions that emphasize **layout**, **flow**, and **expressiveness**.

4.5. Alternative Considered

Custom LoRA Implementation vs. Existing Libraries While several libraries such as `peft` and `diffusers` offer standard LoRA support for Stable Diffusion, we implemented our own LoRA injection mechanism to allow more fine-grained control over where and how LoRA is applied. Existing tools typically focus on patching attention layers only, with fixed defaults and limited configurability. In contrast, our custom setup allowed us to explore alternative strategies, such as injecting LoRA into convolutional layers in residual blocks, varying rank per module, and controlling initialization and update behavior precisely. This flexibility was essential in our setting, where the interplay between spatial layout (handled by attention) and stroke-level texture (handled by convolution) directly affects the aesthetic quality of generated calligraphy. Our approach enabled deeper experimentation with architectural choices and adaptation granularity, which would not have been possible with black-box LoRA toolkits.

5. Experiments

5.1. Hyperparameters and Experimental Setup

We set several key hyperparameters for training our LoRA-enhanced Stable Diffusion model:

- **Batch size:** $[1, 2, 4, 8]$
  We process one training sample per iteration due to memory constraints and the high resolution of images.
- **Learning rate:** $[1 \times 10^{-4}, 1 \times 10^{-5}]$
  This controls the step size during optimization, balancing convergence speed and stability.
- **LoRA rank** r**:** $[2, 4, 8, 16, 32]$
  The rank determines the dimensionality of the low-rank adaptation matrices. Lower ranks reduce model complexity and computation, while higher ranks increase capacity.
- **LoRA scale** $\frac{\alpha}{r}$**:** $[0.5, 1, 2]$
  The scaling factor α normalized by the rank r modulates the contribution of LoRA weights during training.
- **Dropout:** $[0, 0.1, 0.2, 0.5]$
  Dropout rates used to regularize training and prevent overfitting by randomly dropping units during updates.

- **Epochs:** $[10, 20, 50, 100, 200]$
  The number of complete passes over the training dataset.

- **Training steps:** Calculated as

  $$\text{training steps} = \text{epochs} \times \text{training data size}$$

  This represents the total number of iterations performed during training.

5.2. Variations in LoRA Injection Points

To investigate the effect of LoRA weight injection, we experimented with applying LoRA adaptations to different subsets of the model layers:

- **Attention layers only:** Applying LoRA weights exclusively to self-attention modules.

- **All convolution layers + attention layers:** Injection of LoRA weights into every convolution and attention layer, maximizing fine-tuning capacity.

- **Middle convolution layer, up convolution layer, and attention layers:** Targeting the middle and upsampling convolution layers along with attention layers to explore a more focused adaptation.

- **Middle convolution layer and attention layers:** Apply LoRA only to the middle convolution block and attention layers for a balance between specificity and capacity.

This setup helps us understand how different hyperparameters and LoRA injection strategies affect model adaptation, efficiency, and output quality.

5.3. Training Corpus

To explore how LoRA fine-tuning impacts the alignment of visual information with text in different languages, we performed experiments using a consistent set of images that featured English and Chinese text. We applied LoRA adaptations separately to **an English** text corpus and **a Chinese** text corpus, both paired with these same images with the same agumentation methods.

5.4. Evaluation Metrics

5.4.1 Human feedback

We put substantial emphasis on human feedback to evaluate generated calligraphy images due to several factors that are difficult for current computational models to capture:

- **Nuanced Style Variation:** Discerning the finesse and authenticity of a specific calligraphy style often requires human expertise.

- **Text Readability:** Readability can vary significantly based on the style.

- **Aesthetic Appeal:** Calligraphy is an art form, and its evaluation involves subjective aesthetic judgment.

5.4.2 CLIP Evaluation

We utilize the Contrastive Language-Image Pre-training (CLIP) score introduced by Radford [11] as a quantitative measure to assess the semantic similarity between the generated images and their corresponding text prompts.

$$\text{CLIPScore}(x, t) = \frac{\langle f_{\text{img}}(x), f_{\text{text}}(t)\rangle}{\|f_{\text{img}}(x)\| \cdot \|f_{\text{text}}(t)\|} \quad (8)$$

where $f_{\text{img}}(x)$ and $f_{\text{text}}(t)$ are CLIP image embedding and CLIP text embedding.

However, even though CLIP provides a useful automated metric, it has several inherent limitations that can affect its reliability and sensitivity to fine-grained details, e.g. the base CLIP text encoder is not natively well-trained on understanding Chinese prompts. Therefore, though our model gets a 0.18 - 0.25 clip score between prompts and generated calligraphy images, we don't take it as a key factor in evaluating model performance.

5.4.3 FID Evaluation

We use the Fréchet Inception Distance (FID) introduced by Heusel *et al*. [2] to assess the similarity between the distribution of the generated images and the real calligraphy work to quantitatively measure the relevance of generated images.

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \quad (9)$$

where μ_r, μ_g are the mean feature vectors of real and generated images and Σ_r, Σ_g denote their corresponding covariances.

Note that the reliability of our FID computation is restricted by data scarcity, that the amount of real images in our data corpus, 139 Mifu's and 210 Chu Suiliang's, is far less than the recommended minimum 10,000 sample size requirement[2].

5.5. Experiment Results

5.5.1 Experiment Results with Chinese Text Inputs

We conducted extensive experiments with Chinese text prompts to evaluate performance across LoRA configurations. After testing various hyperparameters, the best results were achieved using 100 epochs, a learning rate of 1×10^{-5}, and a dropout rate of 0.1. All results reported use this optimal setup and show significant gains over the baseline, with variations depending on LoRA rank and alpha.

LoRA Rank/Alpha	Base	4/4	8/8	8/16	16/32
Mi Fu FID	532	194	160	204	218
Chu Suiliang FID	519	372	328	315	387
FIDs Sum	1051	566	**488**	549	605

Table 3: FID performance comparison across different LoRA Rank/Alpha configurations and baseline (no fine-tuning). The best (lowest) total FID is **488** at 8/8.

As shown in Table 3, the baseline model (without fine-tuning) produced high FID scores: 532 for Mi Fu and 519 for Chu Suiliang—resulting in a combined score of 1051, indicating a large distribution gap from the target styles.

LoRA fine-tuning improved results across all configurations. The best performance came from rank 8/alpha 8, achieving a combined FID of 488 (a 53.6% improvement), with FID scores of 160 for Mi Fu’ s xingshu and 328 for Chu Suiliang’ s kaishu.

The results highlight style-specific trends: all configurations improved Mi Fu’ s expressive xingshu, with rank 8/alpha 8 performing best. In contrast, Chu Suiliang’ s structured kaishu was more resistant to adaptation, even under optimal settings.

Higher ranks (16/32) did not lead to further gains, suggesting that moderate capacity via LoRA is sufficient. Rank 4/alpha 4 offered a good trade-off between performance and efficiency, delivering competitive results with lower compute costs.

“褚遂良书法”
(Chu Suiliang calligraphy)

“米芾书法”
(Mi Fu calligraphy)

Figure 1: Generated images by best model (rank 8/alpha 8) using Chinese text prompts

As illustrated in Figure 1, the optimal rank 8/alpha 8 model successfully generates high-quality calligraphy when prompted with Chinese text. The left image shows the model’s response to the prompt “褚遂良书法” (Chu Suiliang calligraphy), demonstrating the structured, precise characteristics typical of kaishu style. The right image, generated from the prompt “米芾书法” (Mi Fu calligraphy), exhibits the flowing, expressive brushwork characteristic of xingshu style. These results showcase the model’s ability to differentiate between distinct calligraphic styles and respond appropriately to Chinese character-based prompts.

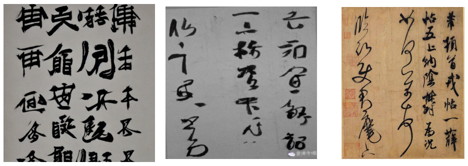
The superior performance of Chinese text inputs compared to baseline generation demonstrates the model’s ability to understand and respond to Chinese character-based prompts after LoRA adaptation. This is particularly noteworthy given CLIP’s inherent limitations with non-English text understanding. The fine-tuning process appears to have enhanced the model’s cross-lingual text-to-image alignment capabilities, enabling more authentic calligraphy generation when prompted with native Chinese descriptions.

These findings highlight the effectiveness of LoRA fine-tuning for domain-specific artistic generation while maintaining computational efficiency. The optimal rank 8/alpha 8 configuration strikes an effective balance between adaptation capacity and overfitting prevention, making it our recommended setting for Chinese calligraphy generation tasks.

5.5.2 Experiment Results with English Text Inputs

As shown in Figure 2, fine-tuning led to substantial improvements in generation quality. Prior to fine-tuning, the model produced abstract or unrecognizable symbols when prompted with English descriptions. When prompted with the Chinese translation (e.g., “米芾行书书法”), the model almost always generated entirely unrelated images, for instance, a rocky landscape with flowers, highlighting its limited understanding of the target domain.

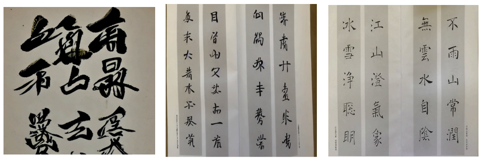
After fine-tuning, the model was able to generate images that were stylistically much closer to the reference calligraphy. In particular, the outputs for prompts such as “米芾” and “褚遂良” demonstrated clear visual alignment with the training samples.



Baseline *Fine-tuned* *Training Image*

Figure 3: Mi Fu xingshu calligraphy

As shown in Figure 3, notably, the lower-left region of the fine-tuned output for “米芾” reflects structural and stylistic features that are consistent with the sample training image, indicating successful adaptation to the target style.



Baseline *Fine-tuned* *Training Image*

Figure 4: chu suiliang kaishu calligraphy

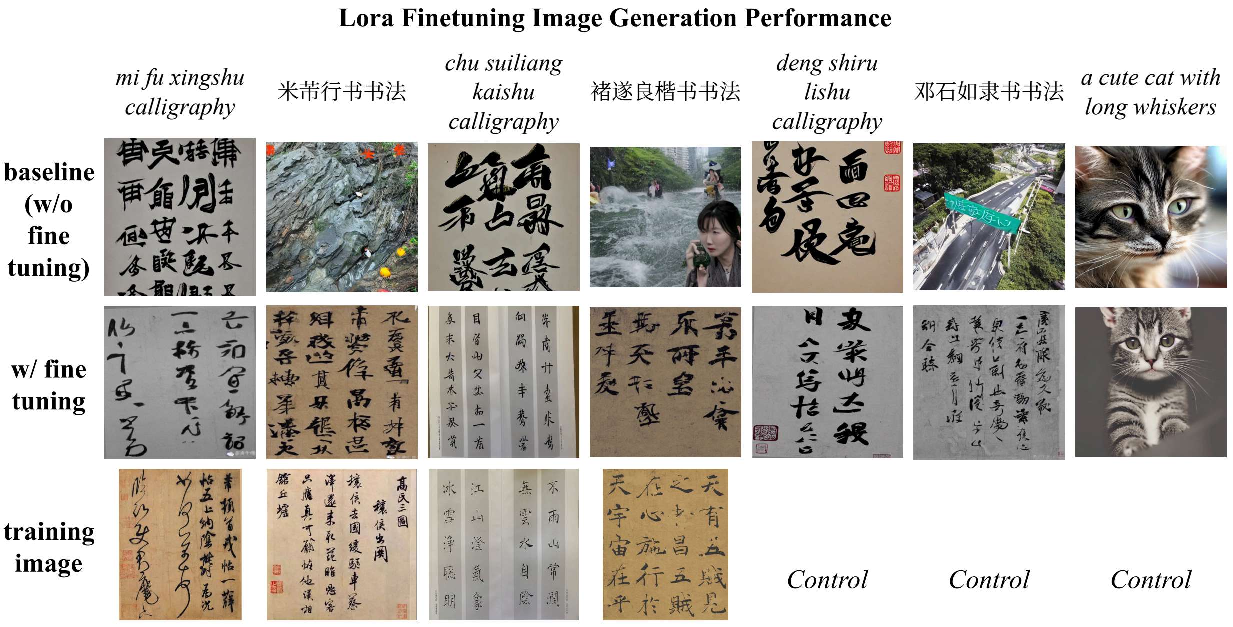

Lora Finetuning Image Generation Performance
mi fu xingshu calligraphy
米芾行书书法
chu suiliang kaishu calligraphy
褚遂良楷书书法
deng shiru lishu calligraphy
邓石如隶书书法
a cute cat with long whiskers
baseline (w/o fine tuning)
w/ fine tuning
training image
Control
Control
Control

Figure 2: Before/after LoRA fine-tuning with selected training images for targeted calligraphy styles.

As shown in Figure 4, even though stylistic alignment showed significant improvement compared to the baseline outputs, limitations remain in character accuracy. For example, in the image generated from the prompt "chu suiliang kaishu," a substantial portion of the generated calligraphy—particularly in the 楷书 (regular script) style—does not correspond to real Chinese characters. This suggests that while LoRA fine-tuning with approximately 348 training samples effectively enabled the model to learn stylistic features, it was insufficient for capturing the full complexity and diversity of Chinese character forms. The results highlight the challenge of simultaneously learning both stylistic and semantic fidelity in high-resolution artistic scripts with limited data.

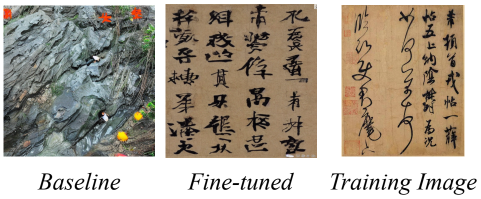

Baseline *Fine-tuned* *Training Image*

Figure 5: 米芾行书

As shown in Figure 5, it is notable that, prior to fine-tuning, the model failed to generate meaningful images when prompted with Chinese character inputs. However, after training exclusively on English text–image pairs, the model demonstrated improved performance even on Chinese character-based prompts. This suggests that the learning acquired through English-based training generalized to some extent across languages. Such cross-lingual transfer indicates that the fine-tuning process enhanced shared representations within the model's multilingual text encoder, despite the absence of explicit alignment between English and Chinese prompts in the original model.

5.6. LoRA Sensitive to Overfitting

During our experiments, we observed a critical overfitting phenomenon when using a learning rate of 1×10^{-4}. The model exhibited severe overfitting that extended beyond the target calligraphy domain, with LoRA-adapted weights becoming so dominant that they interfered with the model's ability to generate images for completely unrelated prompts.

When tested with generic prompts such as "sunset" or "cat", the overfitted model failed to produce coherent images in these domains. Instead, outputs showed calligraphy-like artifacts and features reminiscent of the training data, suggesting that aggressive LoRA updates had corrupted the model's general image generation capabilities by overwriting pretrained knowledge.

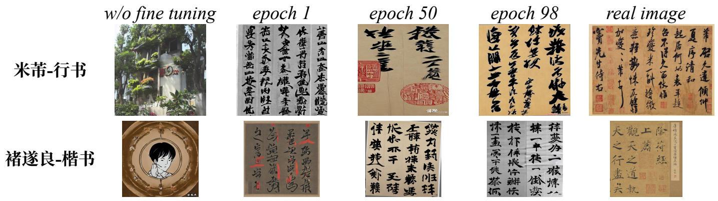


Figure 6: Gradual script style learning

Prompt: “cat” *Prompt: “sunset”*

Figure 7: Generated images for unrelated prompts

Reducing the learning rate to 1×10^{-5} effectively mitigated this issue. The lower learning rate allowed the model to maintain its ability to generate high-quality images for generic prompts while successfully adapting to the calligraphy domain. This finding underscores the delicate balance required in LoRA fine-tuning: adaptations must be strong enough to learn domain-specific features while remaining gentle enough to preserve the model’s general capabilities.

5.7. Rapid Calligraphy, Gradual Script Style

Our experiment shows that the model quickly learns basic calligraphy patterns but struggles with subtle style differences. As shown in Figure 6, it generates calligraphy-like images after just 1 epoch but fails to distinguish between semi-cursive and regular scripts without extended training, highlighting the need for more epochs to capture fine-grained styles.

6. Conclusion

This work demonstrates that LoRA fine-tuning is effective for adapting large models to the complex domain of full-page Chinese calligraphy generation. Using only 350 training samples, the model learned to produce compelling, stylistically diverse images, capturing brush strokes, layout, and artistic flow. The method supports flexible adaptation across art forms, enabling creative stylistic blending.

However, the model struggled to associate specific characters with their calligraphic forms, largely due to CLIP’ s 77-token input limit, which restricts detailed text understanding. Future work could address this by incorporating more powerful language encoders like BERT [1] or T5 [12] to enhance text-image alignment in calligraphy generation.

6.1. Future Work

Addressing Text Input Length Limits A key limitation is the 77-token cap of CLIP’ s text encoder, which restricts the model’ s ability to capture longer calligraphic texts composed of extended phrases or passages. To better model the character-level semantic correspondence, future work could integrate alternative text encoders like [1] or T5 [12], which support longer inputs and richer contextual embeddings. This may alleviate token length constraints and enhance text-to-calligraphy fidelity.

Multilingual Text Understanding Limitation CLIP’ s training on primarily English text-image pairs limits its multilingual understanding, reducing effectiveness on Chinese calligraphy prompts. Future work could incorporate models with stronger multilingual capabilities—such as multilingual BERT [1] or language-specific encoders—to better capture non-English semantics and improve generation quality.

Controlling LoRA Fine-Tuning Influence Future research could explore advanced fine-tuning methods like *dynamic LoRA* [7], which adaptively adjusts adapter parameters during training based on layer importance and input features. This approach allows for more efficient and task-specific optimization, potentially enhancing performance while maintaining computational efficiency.

Contributions

All team members contributed to project through coding and running experimentations.

Acknowledgements

We thank Tiange Xiang for guidance on this project. We utilized the Stable Diffusion implementation from Hugging Face (https://huggingface.co/CompVis/stable-diffusion-v1-4) and built upon the diffusers library for our LoRA fine-tuning experiments.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

[3] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

[5] C. Huang, Q. Liu, B. Y. Lin, T. Pang, C. Du, and M. Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition, 2024.

[6] Q. Liao, G. Xia, and Z. Wang. Calliffusion: Chinese calligraphy generation and style transfer with diffusion modeling, 2023.

[7] X. Liao, C. Wang, S. Zhou, J. Hu, H. Zheng, and J. Gao. Dynamic adaptation of lora fine-tuning for efficient and task-specific optimization of large language models, 2025.

[8] K. Liu, J. Mei, H. Zhang, Y. Zhang, X. Wu, D. Dong, and L. He. Moyun: A diffusion-based model for style-specific chinese calligraphy generation, 2024.

[9] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023.

[10] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, and J. Pfeiffer. Adapters: A unified library for parameter-efficient and modular transfer learning, 2023.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[15] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[16] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning, 2023.