Inheritance and Revitalization: Exploring the Synergy Between AIGC Technologies and Chinese Traditional Culture

Yuhai Zhang, Naye Ji(✉), Xinle Zhu, and Youbing Zhao

Communication University of Zhejiang, Hangzhou 310018, China
jinaye@cuz.edu.cn

Abstract. Diffusion models like Stable Diffusion have made impressive progress in T2I (text-to-image) generation. However, when applied to image generation tasks concerned with Chinese cultural subjects, Stable Diffusion needs to improve the quality of its results. This paper proposes a practical approach to address the challenges of utilizing popular AIGC (AI-Generated Content) technologies and integrating them into a cohesive system, which makes it easier to use Stable Diffusion to create high-quality generated images related to Chinese cultural subjects with direct Chinese prompts. Specifically, with the capabilities of Large Language Models (LLMs), the approach can weave expressive visual descriptions based on initial inputs (scattered words in Chinese, Chinese poems…) and align them in suitable English text prompts for subsequent image generation with Stable Diffusion. Through the parameter-efficient finetuning method called LoRA, Stable Diffusion can effectively learn complex and nuanced concepts of Chinese culture. Additionally, Prompt Engineering plays a role in optimizing inputs, assuring quality and stability, and setting the detailed behavior patterns of LLMs throughout the workflow. This success is attributed to overcoming the constraints of accepting only English prompts and significantly improving the understanding of certain concepts in Chinese culture. The experiments show that our method can produce high-quality images associated with complex and nuanced concepts in Chinese culture by leveraging the fusion of all independent components.

Keywords: Diffusion Models · Large Language Models · Chinese Traditional Culture

1 Introduction

Recently, diffusion models have emerged as promising generative models, particularly for T2I (text-to-image) synthesis. Specifically, large-scale T2I diffusion models such as DALL·E 2 [1], Imagen [2], and Stable Diffusion [3] provide powerful image generation capabilities. It is now possible to generate high-fidelity images based on text prompts. Such powerful T2I image models can be applied in various applications.

The diffusion model possesses remarkable capabilities in visual creativity and multi-style creation. Its application in generating images based on Chinese cultural and artistic

F. Zhao and D. Miao (Eds.): AIGC 2023, CCIS 1946, pp. 24–32, 2024.
https://doi.org/10.1007/978-981-99-7587-7_3

concepts holds significant potential and diverse applications, which can facilitate the dissemination and revitalization of Chinese culture. By enhancing the capabilities of T2I generation models to learn and comprehend artistic mediums like ink painting and meticulous brushwork, shadow puppetry, as well as the intricate patterns and conceptual elements unique to Chinese traditional culture, these valuable cultural heritages can be integrated and innovated with modern features to align with the aesthetics of the general public and obtain new cultural connotations.

However, when applied to image generation tasks associated with Chinese cultural subjects, Stable Diffusion still faces the need to improve the quality of its results due to the constraint of accepting only English prompts and its limited understanding of certain concepts in Chinese culture.

Existing research, such as Taiyi Diffusion [4] and AltDiffusion [5], was developed by training a new Chinese text encoder to address these issues. As a result, even though these models were trained on large datasets, the lack of interaction between the Chinese and CLIP text encoder leads to poor alignments between Chinese, English, and images [6]. Moreover, existing research is often time-consuming and costly. It would be deemed unacceptable to disregard the abundance of valuable open-source resources around Stable Diffusion.

This paper aims to design practicable methods to address these issues within existing Diffusion Model based frameworks. Additionally, it aims to explore the synergy between AIGC technologies and Chinese traditional culture. In this paper, we propose a practical approach to address the challenges associated with the utilization of popular AIGC technologies and align them for integration into a cohesive system.

Specifically, with the capabilities of Large Language Models (LLMs) [7, 10], the approach can weave expressive visual descriptions based on initial inputs (scattered words in Chinese, Chinese poems…) and align them in suitable English text prompts for subsequent image generation with Stable Diffusion. Through the parameter-efficient finetuning method called LoRA [8], Stable Diffusion can effectively learn complex and nuanced concepts of Chinese culture. Additionally, Prompt Engineering plays a role in optimizing input, assuring quality and stability, and setting the detailed behavior patterns of LLMs throughout the workflow.

2 Related Work

2.1 Text-to-Image Synthesis

T2I synthesis has been a subject of considerable interest over an extended period. In the early stage, GAN was a popular choice as the architecture for T2I synthesis models. Recently, large-scale diffusion models have significantly improved the quality and alignment of generated images and replaced GANs in many image-generation tasks [9]. Due to the prosperous open-source community of Stable Diffusion, it is chosen as the default model to implement our T2I generation in this paper.

2.2 Large Language Models

Large language models (LLMs), such as ChatGPT, have attracted enormous attention due to their remarkable performance on various Natural Language Processing (NLP)

tasks. LLMs can produce superior language understanding, generation, interaction, and reasoning capability based on large-scale pre-training on massive text corpora and Reinforcement Learning from Human Feedback (RLHF) [11]. The potent ability of LLMs also drives many emergent research topics to investigate the enormous potential of LLMs further. It brings possibilities for us to build advanced artificial intelligence systems [12]. In this paper, we leverage the capabilities of LLMs to weave expressive English text prompts for Stable Diffusion based on initial inputs.

2.3 LoRA

Low-Rank Adaptation of Large Language Models (LoRA) is a training method that accelerates the training of large models while consuming less memory. It adds pairs of rank-decomposition weight matrices (called update matrices) to existing weights and only trains those newly added weights. With this method to train models, the original pre-trained weights are kept frozen to have multiple lightweight and portable LoRA models for various downstream tasks built on top of them. LoRA was initially proposed for large-language models and demonstrated on transformer blocks, but the technique can also be applied elsewhere. In the case of Stable Diffusion fine-tuning, LoRA can be used for the cross-attention layers that relate the image representations with the prompts that describe them. The method is more effective than other fine-tuning methods for few-shot tasks, such as Textual Inversion [13] and DreamBooth [14].

3 Method

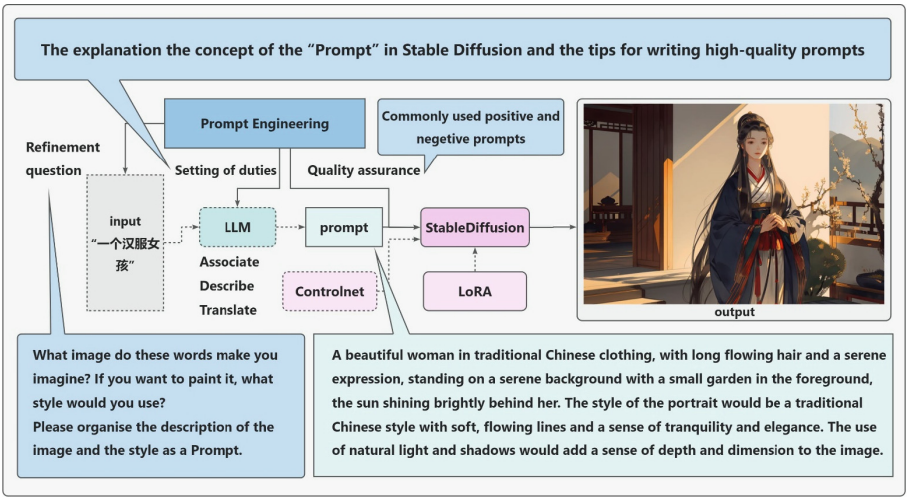

Fig. 1. The workflow of our method.

We propose a practical workflow by leveraging the power of diffusion models, LLMs, prompt engineering, and LoRA. This method provides flexible options for the composition of different structures. Moreover, the type and implementation of components in the workflow are also adjustable. This workflow can produce high-quality images featuring Chinese culture.

The method can be divided into two parts. Composed of LLMs and prompt templates, the first part is responsible for accepting various textual inputs and aligning them into suitable descriptive English text prompts for subsequent image generation with Stable Diffusion. The second part consists mainly of the Stable Diffusion model and our pre-trained LoRA model weights for performing text-image generation tasks, complemented by the ControlNet [15] model to enhance image generation control when needed. The workflow shown in Fig. 1 is introduced in the following subsections in detail.

3.1 Prompt Building

Constructing good picture descriptors is crucial in T2I generation tasks. However, this process is often time-consuming and demands users to possess excellent aesthetic qualities. For image generation with Chinese culture as the theme, our approach involves using Chinese language inputs instead of relying on lengthy lists of English words by capitalizing on the multi-language and multi-profile text understanding capabilities of LLMs, as depicted in Fig. 1. LLMs like ChatGPT or ChatGLM are chosen to aid in constructing the prompts. Cue word templates are designed as guiding prompts for the LLMs, allowing it to generate expressive descriptions tailored to our requirements.

Specifically, we will teach the selected LLMs the concept of the "Prompt" in Stable Diffusion and the tips for writing high-quality prompts. The detailed example prompts designed can be like "*Stable Diffusion, a deep learning text-to-image model that generates images using prompt words to describe the elements to be included or omitted. I introduce the concept of Prompt in the Stable Diffusion algorithm. The prompt here is usually used to describe images and comprises commonly used words. Next, I will explain the steps for generating a prompt, mainly used to describe characters. In generating the prompt, you must describe the character's attributes, themes, appearance, emotions, clothing, posture, perspective, actions, and background using English words or phrases as labels. Then, combine the desired similar prompt words, using a comma as the delimiter, and arrange them from most important to least important. Now, you will serve as the prompt generator.*"

A prompt template is formulated to combine with the user's input, encompassing their information as part of the question. This integration enhances the final returned prompt, providing a more comprehensive step-by-step guide for the model to follow. We can achieve a flexible and adaptable system by incorporating various inputs into the workflow, for instance, by using a poem as input to generate a prompt that closely corresponds to the picture described in the poem. Additionally, even a few Chinese words can yield detailed descriptors, addressing the limitation of only accepting English prompts.

Fig. 2. Prompts generated by LLMs and images generated finally in our method.

3.2 Finetuning with LoRA

While LLMs and prompt engineering can assist in mitigating the challenge of inputting complex English prompts, certain concepts specific to the intricacies of Chinese culture and unique art forms may still require additional enhancement in Stable Diffusion. In such cases, additional training and fine-tuning of the Stable Diffusion model become necessary to enhance its understanding and performance.

By selecting specific elements and art forms from traditional Chinese culture and training them with LoRA, we can enhance the understanding of Chinese cultural concepts in Stable Diffusion, allowing us to achieve the goal of presenting intricate Chinese elements in the T2I task effectively.

3.3 Workflow of the Inference Pipeline

The process is initiated by providing various textual inputs, such as scattered Chinese words or poems, to LLMs, which will respond with descriptive prompts to the user. Depending on the elements the user wishes to showcase in the final image, pre-trained style LoRA weights can be selectively incorporated and combined with the base model

weight while generating the final images. ControlNet can also be employed to further control the final image's structure if needed.

4 Experiments

4.1 Settings

We utilized the gpt-3.5-turbo variant of GPT models and the ChatGLM-6B 1.1 checkpoint as LLMs in the experiments. These models are publicly accessible through the OpenAI API[1] and the Hugging Face repository[2]. Additionally, we incorporated the pre-trained checkpoint model of "TangBohu" for Stable Diffusion.

A WebUI system is developed using standard web development technologies like FastAPI, Spring Boot, and Vue.js to conduct the experiments. The deployment pipeline is built on Diffusers [16] and Transformers [17] libraries. It can integrate all components into a cohesive system with an intuitive user interface, enabling seamless interaction and efficient utilization of the models for image generation with Chinese cultural elements. The component library "Kohya"[3] is utilized for training our LoRA models, and the pre-trained checkpoint model of "TangBohu" for Stable Diffusion serves as the base model for training and inference. The training image datasets for our models were collected online by ourselves. For image captioning, we first use WD1.4 ViT Tagger[4] for automatic labeling and then manually make adjustments.

4.2 Results

We present several demonstrations in Fig. 1, Fig. 2, Fig. 3, Fig. 4, and Fig. 5. Figure 3 illustrates the outcomes of the artistic creation function using the Stable Diffusion combined with multiple models we trained. These models encompass elements in Chinese traditional culture like Chinese ink painting and meticulous brushwork, shadow puppetry, as well as the intricate patterns of Dunhuang and Blue-and-White.

Fig. 1 and Fig. 2 represent prompts generated by LLMs and final images generated with our method. The prompts in Fig. 2 are generated by gpt-3.5-turbo while those in Fig. 1 are generated by ChatGLM-6B. In addition, only images in Fig. 2 were directly generated by the basic model without using any models we trained, so they can be compared and referenced with images in other figures. Figures 4 and 5 represent case studies in our experiments. Figure 4 showcases pattern images generated by the trained pattern model, along with the results obtained from combining the "ControlNet" depth model for interior design rendering. Meanwhile, Fig. 5 demonstrates the outcomes of various pre-trained models combined with the "ControlNet" OpenPose model (detects and copies human poses without copying other details), focusing on traditional Chinese costume design.

[1] https://platform.openai.com/.
[2] https://huggingface.co/THUDM/chatglm-6b.
[3] https://github.com/bmaltais/kohya_ss.
[4] https://github.com/picobyte/stable-diffusion-webui-wd14-tagger.

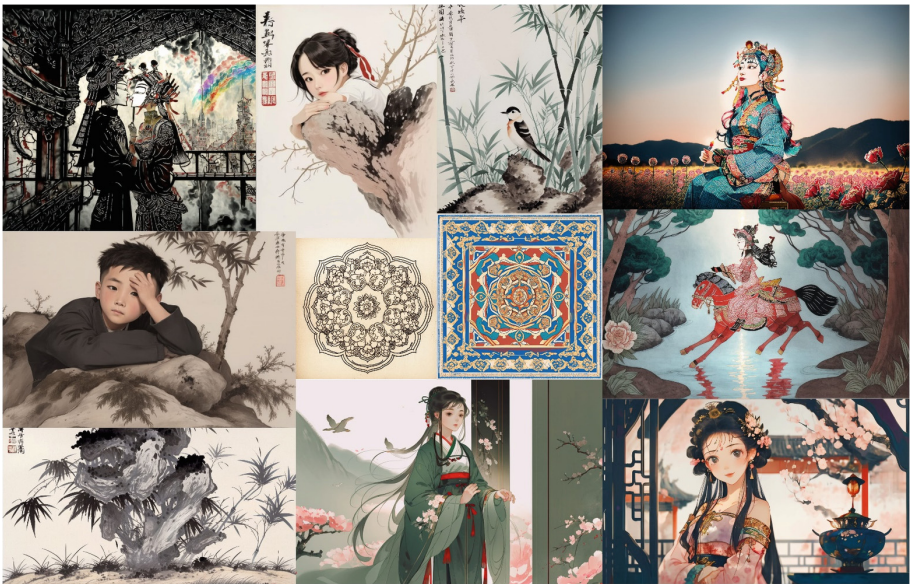

Fig. 3. Images featuring Chinese culture generated by our method.

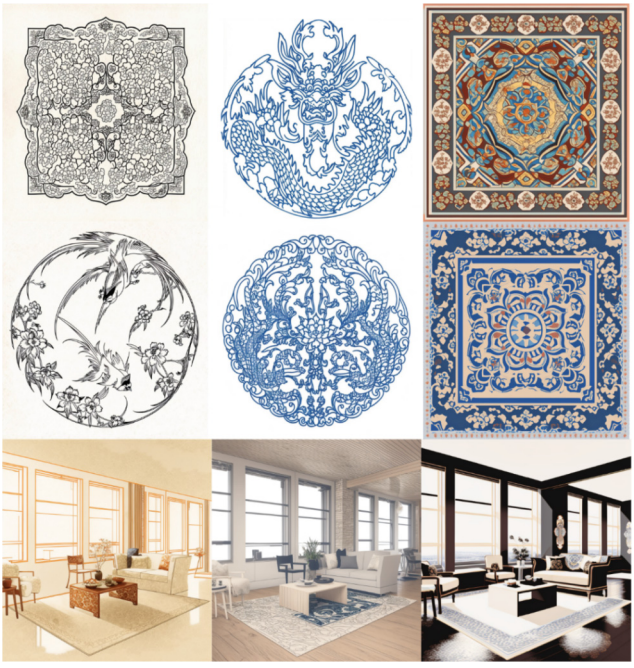

Fig. 4. Images generated with our method.

Fig. 5. Images generated with our method.

5 Conclusions

This paper presents a practical approach to integrating popular AIGC technologies into a unified system while also exploring the potential synergy between AIGC technologies and Chinese traditional culture. Specifically, with the capabilities of LLMs, the approach can weave expressive visual descriptions based on initial inputs and align them in suitable English text prompts for subsequent image generation with Stable Diffusion. Through the fine-tuning method called LoRA, Stable Diffusion can effectively learn complex and nuanced concepts of Chinese culture. The experimental results demonstrate that our approach can generate high-quality results pertaining to Chinese cultural subjects. This success is attributed to overcoming the constraints of accepting only English prompts and significantly improving the understanding of certain concepts in Chinese culture.

Acknowledgments. This paper is partially supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2023C01212) and the Public Welfare Technology Application Research Project of Zhejiang (No. LGF22F020008).

References

1. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latent. arXiv preprint arXiv:2204.06125 (2022)

2. Saharia, C., et al.: Others: photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural. Inf. Process. Syst. **35**, 36479–36494 (2022)
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695 (2022)
4. Wang, J., et al.: Fengshenbang 1.0: being the foundation of Chinese cognitive intelligence. CoRR. abs/2209.02970 (2022)
5. Chen, Z., Liu, G., Zhang, B.-W., Ye, F., Yang, Q., Wu, L.: AltCLIP: altering the language encoder in CLIP for extended language capabilities (2022). https://doi.org/10.48550/ARXIV.2211.06679
6. Saxon, M., Wang, W.Y.: Multilingual conceptual coverage in text-to-image models. arXiv preprint arXiv:2306.01735. (2023)
7. Du, Z., et al.: Glm: general language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360. (2021)
8. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. (2021)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
10. Brown, T., et al.: Others: language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
11. Ouyang, L., et al.: Others: training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
12. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: solving AI tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580. (2023)
13. Gal, R., et al.: An image is worth one word: personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618. (2022)
14. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510 (2023)
15. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543. (2023)
16. Platen von, P., et al.: Diffusers: state-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)
17. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online (2020)