Oracle Bone Character Generation with Diffusion Models

Xiaoxuan Xie[1,a)] Xusheng Du[1] Minhao Li[1] Haoran Xie[1]

Abstract: Generating is a challenging task due to their scarcity and unique pictographic nature. In this research, we propose a novel approach to generate oracle bone character-style images corresponding to the given image input using diffusion models. Unlike existing datasets of oracle bone characters, they typically store paired characters with modern Chinese characters. We first construct the dataset that aligns oracle bone characters, text prompts, and object images. We then train a stable diffusion model on this dataset. By inputting images of different objects combined with corresponding text prompts, the model generates the corresponding images in the style of oracle bone characters. Additionally, we integrate an optimization module to refine the initial results, ensuring they better conform to the original structure and norms of oracle bone characters. The qualitative evaluation demonstrates that our model excels in generating oracle bone character-style images that are both stylistically and semantically consistent.

Keywords: Oracle bone characters, Image generation, Stable diffusion model

1. Introduction

Oracle bone character (OBC) is one of the oldest well-known writing systems in the world, dating back over 3,500 years (Fig. 1). However, their scarcity and unique pictographic nature present significant challenges in archeology and visual computation. Despite the discovery of thousands of oracle bone fragments, only 4,500 unique characters are identified successfully, with two-thirds remaining unidentified. Unlike modern language characters, OBCs lack a standardized coding system for digitization and text storage, resulting in their preservation mainly in image form and the absence of a comprehensive corpus. Additionally, the pictographic style requires extensive knowledge from multiple disciplines, such as archaeology, history, and philology, and there are few experts capable of recognizing and annotating these characters. Inspired by the unique styles of oracle bone characters, we propose generating sufficient and controllable OBC-style images using image generation models, leveraging the potential pictographic interpretations of the characters, as shown in Fig. 2.

Nowadays, artificial intelligence technology has been explored extensively to understand and study this ancient language. Researchers have utilized GAN models to train directly on existing OBCs [5] or to extend writing styles through handwritten OBCs [6]. While these GAN-based approaches can generate a sufficient number of OBCs to improve recognition performance, they often lack controllability and suffer from unstable training issues. In comparison, diffusion models (DM), such as Stable Diffusion [11] and ControlNet [15], have demonstrated significant potential in synthesizing high-quality images due to their stable learning objectives and controllable text prompts [13]. However, these researches focus on the recognition and detection of deciphered OBCs, lacking text prompts for generated content and image interpretations of OBCs.

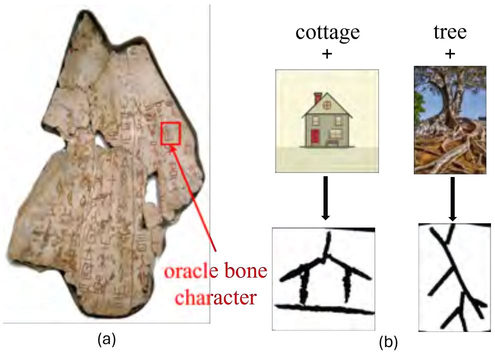

Fig. 1 An oracle bone fragment photographed by BabelStone at the National Museum of China[*2](a). This work aims to combine text and images to generate OBC-style characters (b).

In this research, we propose a novel approach to generating OBC-style images from given image inputs using a diffusion model. We first construct a dataset that aligns OBCs with text prompts and object images. We then train a ControlNet model on this dataset. By inputting images of various objects along with corresponding text prompts, the model generates the corresponding OBC-style images. Additionally, we integrate an optimization module to refine the initial results, ensuring they better conform to the original structure and style of OBCs. Qualitative evaluation indicates that our model effectively produces OBC-style images that are both stylistically and semantically consistent.

The main contributions of this work are listed as follows:

- We constructed a dataset with images of original oracle bone

[1] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923–1211, Japan
[a)] s2310069@jaist.ac.jp

[*2] File:Shang dynasty inscribed scapula.jpg. (2021, October 31). Wikimedia Commons. Retrieved 09:08, June 6, 2024 from `https://commons.wikimedia.org/w/index.php?title=File:Shang_dynasty_inscribed_scapula.jpg&oldid=604424222`.

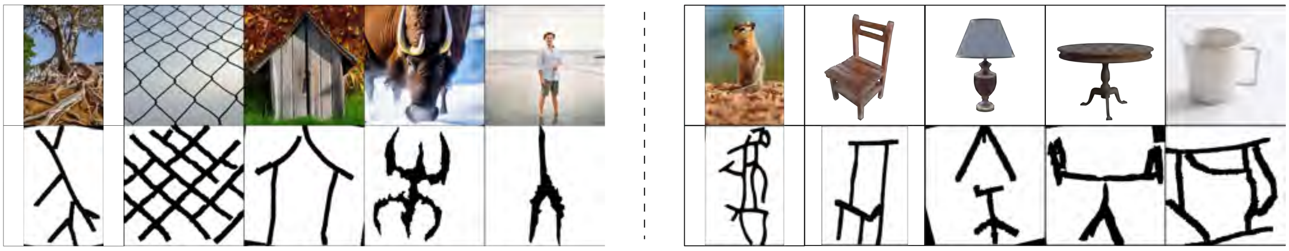

Fig. 2 OBC-style images: The first row shows input images for generation, and the second row shows images generated by our method. The left side shows reconstructed images from the original dataset, while the right side shows generated images of non-existing characters.

characters, corresponding real-world images, and descriptive textual prompts.

- We proposed a novel two-stage approach for generating oracle bone character-style images from given images and text input, ensuring the characters are stylistically accurate and semantically meaningful.

2. Related Works

2.1 Oracle Bone Characters

Through high-precision scanning of OBCs obtained from archaeological excavations, significant and promising results have been achieved in computer-based research on OBCs. Guan et al.[3] created the EVOBC dataset for studying the evolution of OBCs in Ancient Chinese literature. This dataset includes scanned images of OBCs from various sources, greatly aiding in identifying the corresponding modern Chinese characters. In terms of feature extraction for OBCs, Li et al. [6]used the Stable Diffusion model [11] to recognize features of hand-drawn OBCs, transforming scanned OBCs from ancient texts into a style that resembles hand-drawn illustrations. Du et al. [2] employed deep self-supervised learning to extract feature information from OBCs. This method can simulate simple lines in the style of OBCs but performs less effectively when handling images with extensive RGB information and distracting elements. Due to the ancient origin of OBC, the limited number of unearthed samples, and the uneven distribution of datasets, Yue et al. [14] proposed a new deep learning model called C-A Net. This model uses modified Generative Adversarial Networks (GANs) for dynamic dataset augmentation, effectively addressing the issue of overfitting caused by imbalanced datasets.

2.2 Conditional Image Generation

Typical conditional image generation models include cGAN [7] and diffusion models [13], which can dynamically adjust the output based on specific requirements. Earlier work [1], [8], [13] primarily involved extracting relevant features from the RGB regions of images and generating related images in a similar style without any prompts based on these features. The Latent Diffusion Model [11] uses latent space, optimizing computational complexity compared to traditional Diffusion Models. Additionally, Conditioning Mechanisms enhance the U-Net [12] through cross-attention mechanisms, making conditional image generation more flexible.

Recent conditional image generation can leverage the strengths of StyleGAN for sketch-based latent space exploration to achieve high-quality synthesis from incomplete sketches [4]. The Stable Diffusion Model[11], which integrates Latent Diffusion Model functionalities. By introducing CLIP (Contrastive Language-Image Pre-Training)[10], it generates images with relevant features from prompts, addressing the resource consumption and accuracy limitations of building diffusion models directly in high-dimensional feature spaces. The generated models are more accurate, support higher image resolutions, and better meet user needs than the standalone Diffusion Model. However, Stable Diffusion requires extensive keyword combinations to present a relatively good representation. When generating more abstract images, there might be a probability of distortion and deformation in some areas. As an extension of Stable Diffusion, ControlNet[15] intervenes in the denoising process of images by adjusting the neural network, enabling it to have the chance to generate stable images in a specified direction under the guidance of prompts. The conditional generation using stable diffusion can be adopted in various domains including fashion design[17], scene[16] and facial image generations[9].

3. Method

In this section, we introduce our proposed two-stage approach for generating OBC-style images, as shown in Fig. 3. In Section 3.1, we present ControlNet, which serves as the primary reference for our method. During the generation stage, OBC-style images are created based on user input images and prompts (Section 3.2). In the optimization phase, style optimization, edge refinement, and resolution enhancement are applied to obtain the final refined results (Section 3.3).

3.1 ControlNet

Our approach leverages ControlNet[15], which extends the capabilities of base diffusion models by enabling conditional image generation. ControlNet generates target results by incorporating control information, such as text descriptions and reference images into the generative process. Formally, the pre-trained stable diffusion model, denoted as $N(\cdot;\Theta)$ with parameters Θ, takes an input feature map x and produces the corresponding image feature maps y, which can be expressed as:

$$y = N(x;\Theta) \tag{1}$$

While ControlNet freezes the initial weights Θ of the pre-trained model and creates a copy of the encoder and middle blocks. This trainable copy, Θ_c, is specifically responsible for

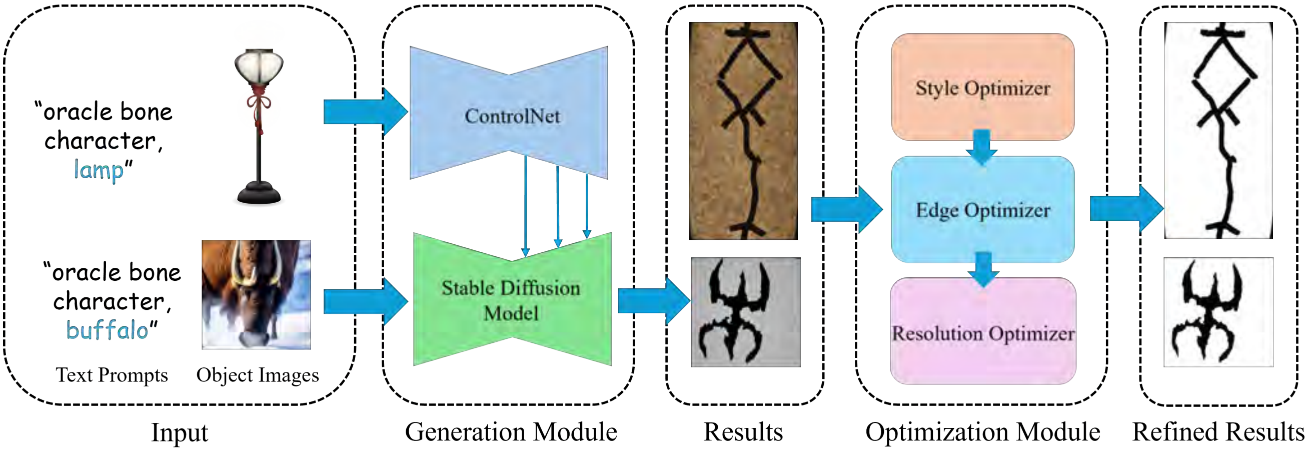

Fig. 3 The framework of our proposed method. First, textual prompts and object images are used as inputs to generate initial results through the generation module. Next, these initial results are passed through style, edge, and resolution optimizers to obtain refined results.

processing the control conditions c. The trainable copy is then integrated back into the pre-trained model's decoder using two zero-weight convolutional layers, $N(\cdot;\Theta_{z1})$ and $N(\cdot;\Theta_{z2})$. These zero-weight layers initially have minimal impact on the generative process but gradually learn to modulate the output as training progresses. The final output of the ControlNet block, y_c, can be expressed as:

$$y_c = F(x;\Theta) + Z\left(N\left(x + Z(c;\Theta_{z1});\Theta_c\right);\Theta_{z2}\right) \qquad (2)$$

Through this process of controlling conditions in ControlNet, we can effectively leverage the pre-trained stable diffusion model's generative capabilities to generate images that adhere to the specified OBC-style conditions.

3.2 Generation Module

In the generation phase, we utilize the fine-tuned image-to-OBC ControlNet as an additional control and use a stable diffusion model to generate OBC-style images. The text prompts and object images are used as conditional inputs and are converted by the encoder into latent representations τ_θ. Simultaneously, the object images are encoded into latent features c_i. During the diffusion process, random latent noise (z_t) is iteratively backpropagated through the U-Net [12] network to obtain (z_{t-1}). The U-Net network can be represented as:

$$\epsilon_\theta(z_t, t, \tau_\theta) \qquad (3)$$

where t denotes each time step. Simultaneously, c_i is input into ControlNet to obtain control features c_f. By using c_f as an additional control condition, we can obtain the updated representation:

$$\epsilon_\theta(z_t, t, c_f, \tau_\theta) \qquad (4)$$

By iterating the U-Net, we obtain $z_{t-1}, z_{t-2}, \ldots, z_0$. The z_0 is decoded to generate the final OBC-style image as the initial result.

3.3 Optimization Module

After generating the initial result images, we perform a unified optimization process to obtain the final results. This phase mainly involves three parts: style optimization, edge refinement, and character resolution enhancement.

3.3.1 Style Optimization

In this part, our objective is to accurately distinguish between the OBCs region and the background region for further image processing. OBCs typically appear as deep black, while the background region may contain various colors. Given that the OBCs in the generated images are nearly black, its pixel values generally do not exceed 50.

To achieve this distinction, we employ pixel value detection and binarization. Specifically, we classify regions with pixel values greater than 50 as background regions, and those with pixel values less than 50 as OBCs regions. For ease of subsequent processing, these pixel values are set to 0 (black) for the script and 255 (white) for the background. This binarization process can be described by the following equation:

$$P'(x,y) = \begin{cases} 0, & \text{if } P(x,y) < 50 \\ 255, & \text{if } P(x,y) \geq 50 \end{cases} \qquad (5)$$

where $P(x,y)$ represents the pixel value at position (x,y) in the original image, and $P'(x,y)$ represents the pixel value after binarization. This threshold-based binarization method is not only simple and efficient but also meets the needs of practical applications. By using this method, we can effectively separate the OBCs region from the background, laying the foundation for subsequent image analysis and processing.

3.3.2 Edge Refinement

For edge refinement, we employ a Gaussian smoothing filter to refine the edges of the generated OBC-style images, effectively reducing noise while preserving detail. The Gaussian smoothing filter operates by convolving an image with a Gaussian function, which helps to achieve clearer and sharper edges.

Initially, a Gaussian kernel is created based on the Gaussian function, with the kernel size and standard deviation (σ) determining the degree of smoothing. The 2D Gaussian function is defined as follows:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (6)$$

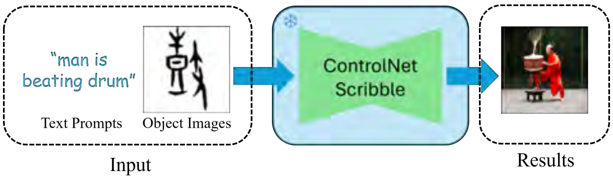

Fig. 4 With the original OBCs data and the OBCs meaning, aligns images corresponding to the OBCs are generated by the pre-trained ControlNet scribble model.

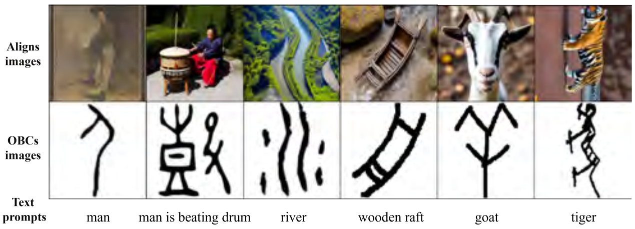

Fig. 5 The dataset includes textual prompts, images of original OBCs, and aligns images. The dataset contains 30 different original OBCs images, and 5 sets of data per OBCs, totaling 150 data pairs.

where $G(x, y)$ represents the value of the Gaussian function at point (x, y), the variables x and y are the coordinates of a point in the 2D space relative to the center of the Gaussian kernel. The image is then convolved with the Gaussian kernel. Convolution involves sliding the kernel across each pixel in the image, multiplying the kernel values with the corresponding pixel values in the neighborhood, and summing these products to obtain the new pixel value. The specific formula is defined as follows:

$$I'(i, j) = \sum_{k=-m}^{m} \sum_{l=-m}^{l} G(k, l) \cdot I(i + k, j + l) \qquad (7)$$

where $I'(i, j)$ is the new value of the pixel at position (i, j) after applying the Gaussian filter, $I(i + k, j + l)$ is the value of the pixel at position $(i + k, j + l)$ in the original image, and $G(k, l)$ is the value of the Gaussian kernel at position $(k, l).m$ is the radius of the Gaussian kernel. The application of the Gaussian smoothing filter preserves edge details while effectively reducing noise. To further optimize the results, binarization is applied to the smoothed image, enhancing the definition and sharpness of the OBCs edges. This method significantly improves the visual quality and ensures stylistic and structural consistency with authentic OBCs, thereby enhancing character recognition performance and the overall quality of the research outcomes.

3.3.3 Resolution Enhancement

For resolution enhancement, we apply bilateral filtering to enhance the image resolution of the generated OBC-style images. Bilateral filtering is a non-linear, edge-preserving, and noise-reducing filter. Unlike traditional Gaussian filters, bilateral filters consider both spatial distances and intensity differences, effectively preserving edges while smoothing the image. The output of a bilateral filter is calculated as follows:

$$I'(x) = \frac{1}{W_p} \sum_{x_i \in S} I(x_i) \cdot f_r(|I(x_i) - I(x)|) \cdot f_s(|x_i - x|) \qquad (8)$$

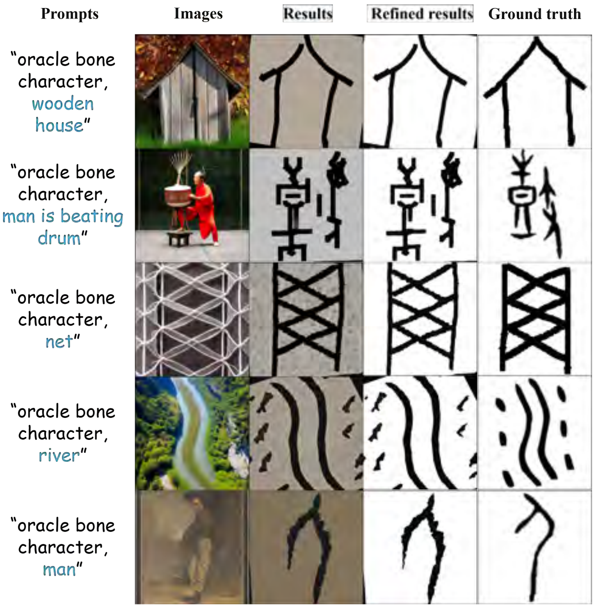
Prompts
Images
Results
Refined results
Ground truth
"oracle bone character, wooden house"
"oracle bone character, man is beating drum"
"oracle bone character, net"
"oracle bone character, river"
"oracle bone character, man"

Fig. 6 Generation results.

where $I(x)$ represents the original intensity of pixel x, $I'(x)$ the filtered intensity, x_i the neighboring pixels, and S the spatial domain. The range kernel f_r smooths intensity differences, while the spatial kernel f_s smooths spatial differences. W_p is a normalization factor ensuring the sum of the weights to 1.

We apply the bilateral filter to the generated images. This step reduces noise and preserves edges by considering both spatial and intensity information. After bilateral filtering, bicubic interpolation is used to upscale the images to the desired resolution. This process increases the pixel count, producing higher-resolution images while maintaining sharpness and detail. By incorporating bilateral filtering, we significantly enhance the resolution and visual quality of the generated OBC-style images, ensuring they are detailed and clear, thus facilitating better analysis and application.

4. Experiment and Results

We validate the similarity of the model-generated images to the style of the OBCs and input images through qualitative experiments. In section 4.1, we present the implementation details. In section 4.2, we show the results of the qualitative evaluation.

4.1 Implementation Details

In the dataset creation phase, we first select the most original OBCs dataset OBC from the dataset of EVOBC[3], and select 30 groups of characters from it, with 5 images for each group of characters, totaling 150 images of original OBCs. After that, the meanings expressed by all the OBCs are used as prompts to obtain the aligned images by pre-trained ControlNet model[15], as shown in Fig. 4. Finally, we got 150 data pairs as shown in Fig. 5.

During the training phase, we fine-tuned the v2-1-512-ema-pruned model of Stable Diffusion using ControlNet. Given the limited amount of data, we employed various data augmentation techniques, including random horizontal flip, random vertical flip, random rotation (in multiples of 45°), random cropping and resizing, as well as random adjustments to brightness, contrast, saturation, and tone. The model was trained for 1500 epochs with a batch size of 4 on a single NVIDIA A100 GPU.

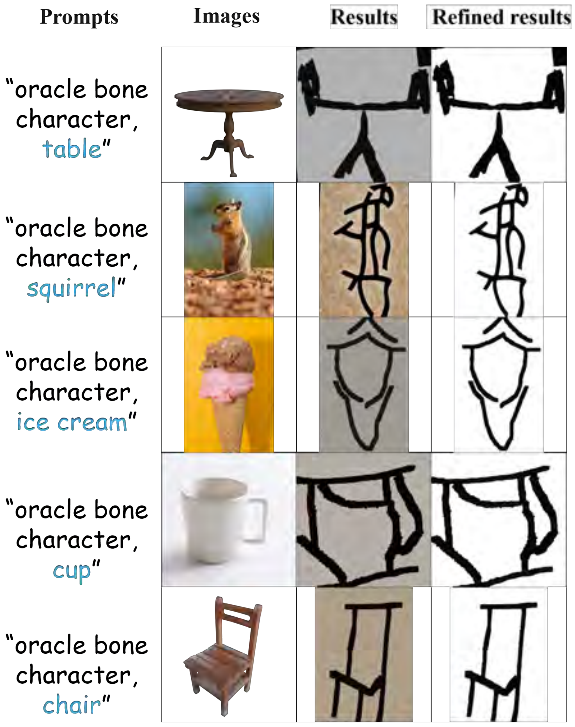
Prompts
Images
Results
Refined results
"oracle bone character, table"
"oracle bone character, squirrel"
"oracle bone character, ice cream"
"oracle bone character, cup"
"oracle bone character, chair"

Fig. 7 Results of generated images of objects not present in the original dataset, including chairs, cups, ice cream, squirrels, and tables.

4.2 Qualitative Evaluation

We evaluated our model from three aspects: reconstruction of existing items from the original dataset, generation of new items not present in the original dataset, and generation of different styles for the same items.

Reconstruction of Existing Characters. We assessed the model's ability to faithfully reproduce existing items from the training dataset. We selected five diverse samples encompassing buildings, simple scenes, objects, landscapes, and human images. These samples were processed through our network to generate refined outputs. As shown in Fig. 6, the results exhibit high fidelity to the ground truth, demonstrating the model's effectiveness in preserving semantic information and adhering to the stylistic conventions of OBCs.

Generation of Non-existent Characters. We evaluated the model's ability to generate images for objects not present in the original dataset, such as chairs, cups, ice cream, squirrels, and tables. The generated images, illustrated in Fig. 7, successfully capture the distinctive style of OBCs, highlighting the model's proficiency in maintaining stylistic consistency even for unseen objects.

Generation Diversity. We explored the model's ability to generate diverse stylistic variations within a single category. Using the same text prompt ("oracle bone character, lamp"), we fed the model with images representing five different styles of table lamps. As shown in Fig. 8, the model effectively captured both the semantic content (the concept of a lamp) and the stylistic variations present in the input images, while maintaining the overall aesthetic of OBCs.

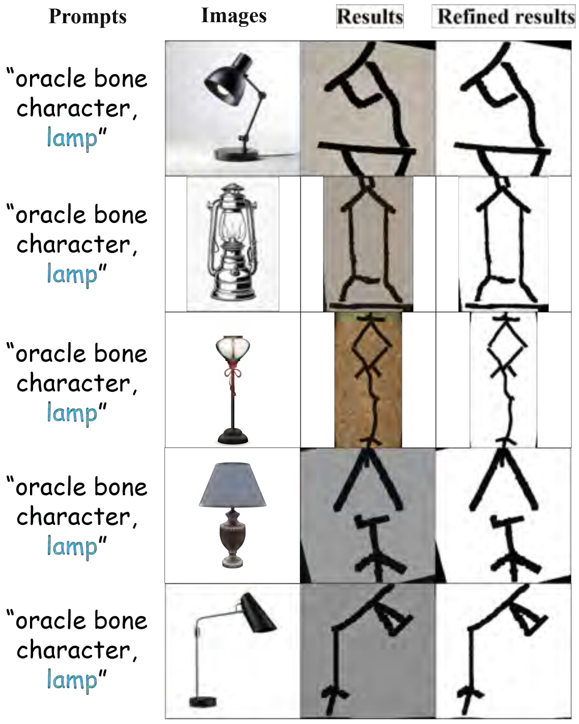
Prompts
Images
Results
Refined results
"oracle bone character, lamp"
"oracle bone character, lamp"
"oracle bone character, lamp"
"oracle bone character, lamp"
"oracle bone character, lamp"

Fig. 8 Examples of generated images showing stylistic variation within a single category. These results capture both semantic content and stylistic differences while maintaining the overall aesthetic.

5. Conclusion

In this work, we proposed image generation approach to generate images in the style of oracle bone characters using diffusion models. By constructing a dataset that aligns OBCs, text prompts, and object images, we trained a stable diffusion model capable of producing stylistically consistent outputs from diverse input images and text prompts. Our model integrates an optimization module that refines the initial results, ensuring better conformity to the original structure and norms of OBCs. Both qualitative and quantitative analyses demonstrate that our method achieves excellent results in generating semantically consistent OBC-style images.

However, our method has certain limitations. Firstly, the model struggles with images that have complex backgrounds or distinct lines, resulting in less effective generation outcomes for these cases. Secondly, due to the rotational data augmentation applied during training, some generated images exhibit small-angle rotational artifacts, such as black edges, which affect the overall visual quality. Despite these limitations, our approach represents a significant advancement in generating OBC-style images. It provides a robust framework for further research and development in this OBC-related field.

References

[1] Choi, J., Kim, S., Jeong, Y., Gwon, Y. and Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models, *arXiv preprint arXiv:2108.02938* (2021).

[2] Du, B., Liu, G. and Ge, W.: Deep Self-Supervised Learning for Oracle Bone Inscriptions Features Representation, *2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, IEEE, pp. 7–11 (2021).

[3] Guan, H., Wan, J., Liu, Y., Wang, P., Zhang, K., Kuang, Z., Wang, X.,

Bai, X. and Jin, L.: An open dataset for the evolution of oracle bone characters: EVOBC, *arXiv preprint arXiv:2401.12467* (2024).

[4] Huang, Z., Xie, H., Fukusato, T. and Miyata, K.: AniFaceDrawing: Anime Portrait Exploration during Your Sketching, *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3588432.3591548 (2023).

[5] Li, J., Wang, Q.-F., Huang, K., Yang, X., Zhang, R. and Goulermas, J. Y.: Towards better long-tailed oracle character recognition with adversarial data augmentation, *Pattern Recognition*, Vol. 140, p. 109534 (2023).

[6] Li, J., Wang, Q.-F., Huang, K., Zhang, R. and Wang, S.: Diff-Oracle: Diffusion Model for Oracle Character Generation with Controllable Styles and Contents (2023).

[7] Mirza, M. and Osindero, S.: Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* (2014).

[8] Nichol, A. Q. and Dhariwal, P.: Improved denoising diffusion probabilistic models, *International conference on machine learning*, PMLR, pp. 8162–8171 (2021).

[9] Peng, Y., Zhao, C., Xie, H., Fukusato, T. and Miyata, K.: Sketch-Guided Latent Diffusion Model for High-Fidelity Face Image Synthesis, *IEEE Access*, Vol. 12, pp. 5770–5780 (online), DOI: 10.1109/ACCESS.2023.3346408 (2024).

[10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision, *International conference on machine learning*, PMLR, pp. 8748–8763 (2021).

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

[12] Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, pp. 234–241 (2015).

[13] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics, *International Conference on Machine Learning*, PMLR, pp. 2256–2265 (2015).

[14] Yue, X., Li, H., Fujikawa, Y. and Meng, L.: Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition, *ACM Journal on Computing and Cultural Heritage*, Vol. 15, No. 4, pp. 1–20 (2022).

[15] Zhang, L. and Agrawala, M.: Adding conditional control to text-to-image diffusion models, *arXiv preprint arXiv:2302.05543* (2023).

[16] Zhang, T. and Xie, H.: Sketch-Guided Text-to-Image Generation with Spatial Control, *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*, pp. 153–159 (online), DOI: 10.1109/CGIP62525.2024.00035 (2024).

[17] Zhang, Y., Zhang, T. and Xie, H.: TexControl: Sketch-Based Two-Stage Fashion Image Generation Using Diffusion Model (2024).