

Received 27 August 2024, accepted 17 September 2024, date of publication 20 September 2024,
date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3464562

RESEARCH ARTICLE

GlyphGenius: Unleashing the Potential of AIGC in Chinese Character Learning

LINHUI WANG¹, YI LOU¹, XINYUE LI¹, YUXUAN XIANG², TIANYI JIANG²,
YIYING CHE¹, AND CHEN YE^{1,3}, (Member, IEEE)

¹College of Electronic and Information Engineering, Tongji University, Shanghai 200092, China

²College of Design and Innovation, Tongji University, Shanghai 200092, China

³Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai 200092, China

Corresponding author: Chen Ye (yechen@tongji.edu.cn)

ABSTRACT Unlike phonetic writing systems such as English, Chinese characters, as ideographic symbols, combine sound, form, and meaning. Therefore, learning and mastering Chinese characters inevitably involve understanding their etymology and semantics represented by their shapes. However, for the sake of writing convenience, Chinese characters have undergone a long process of evolution, gradually reducing their pictographic components. This evolution poses greater challenges for non-native learners unfamiliar with the structure of square Chinese characters. In this paper, we propose a novel approach to assist in Chinese character learning. Considering the unique visual features deeply rooted in the cultural origins of Chinese characters and the ability of Artificial Intelligence Generated Content (AIGC) to generate images without requiring user expertise, we utilize the AIGC model to redraw Chinese character components based on their inherent meanings. Through this visual transformation, users can intuitively grasp the semantics of Chinese characters, opening up new avenues for Chinese character learning. The Guess-Meaning experiment reveals that learners with less than one year of experience scored an 12.76% higher accuracy in recognizing the meaning of characters that had been redrawn, as compared to the original characters. During system usability testing, users reported an average satisfaction rating of 4.24 out of 5 points. The major limitations of the present study are that the current system still relies on human understanding of Chinese characters for redrawn prompts, and not all Chinese characters have corresponding pictorial meanings. The system and repository are now accessible via the link <https://scroll.ihanzi.net> and <https://github.com/BlossomsGarden/Glyph-Genius>.

INDEX TERMS Glyph visualization, human-computer interaction, stable diffusion, visualization in education.

I. INTRODUCTION

Characters are the basis for maintaining the cultural identity and national centripetal force of a country or region, nurturing public morals, and passing on ideas and civilisations, as well as an important carrier for inspiring individual poetry, aesthetic intellect and free spirit. Chinese is one of the oldest and also the most widely used scripts in the world. In contrast to English and other phonogram, Chinese characters, as logograms, carry both visual semantics (form) and textual semantics (meaning) to some extent. The characteristics of

Chinese character in terms of visual flexibility and semantic metaphorical sets itself apart from other language systems. As is shown in figure 1, the meaning of pictograms in Chinese can be seen intuitively in both components and characters-words. Therefore, learning and mastering Chinese characters inevitably involve understanding the root and origin of the characters.

In previous studies related to Chinese character learning, many have focused on the learning patterns, exploring more rational learning methods [1], [2]. Other researchers have been dedicated to creating richer interactive media [3], [4], such as speech recognition modules [5] and song generation module [6]. However, these are just parts of the issue.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia¹.

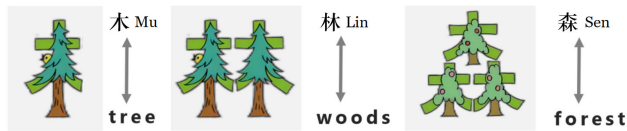


FIGURE 1. Chinese character “木”“林”“森” are typical pictograph. “林” means some trees, similar to “woods” in its textual semantics. As for its visual semantics, it’s easy to recognize “林” is a combination of “木”, which means a single tree, indicating that there are a few trees. “森” is composed of three “木” in its structure, indicating that there are many trees.

Various innovations in interactive devices enable users to interact in a more convenient manner, but in aiding users’ memorization of Chinese characters, researchers often still need to manually input explanatory videos or learning materials on character construction. The manual uploading of learning materials itself poses a significant barrier to system setup and maintenance. The Latent Diffusion Model (LDM) [7] has successfully commercialized and released the pre-trained model stable-diffusion. However, samples in figure 2 (a) show different effect of generation between Chinese characters and English words. The pre-trained model’s ability in generating normally structured Chinese characters is limited, resulting in artistic but distorted representations of the characters’ skeletons. Other Models like DALL-E-2 [8] also encounter similar issues, displaying a greater proficiency in handling Western fonts with simple structure compared to Chinese characters. Therefore, we try to better enable generative models to learn the stylistic features of the writing and skeleton of Chinese characters. Then we can deploy the model to Chinese character learning, which will help avoid the laborious task of creating and uploading various learning materials in Chinese character learning media.

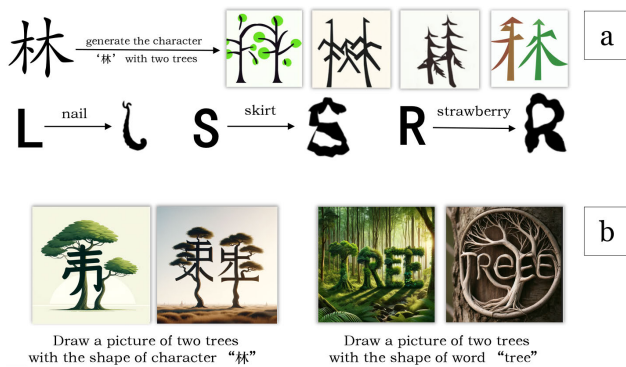


FIGURE 2. (a) Utilizing the open-source Stable Diffusion pre-trained model [9] directly, demonstrate the effect of drawing two trees based on the Chinese character “林” and the letters L, S, and R. (b) Using DALL-E-2 model to generate two images based on the Chinese character “林” and the word “tree” respectively. We access DALL-E 2 as a generation tool that is integrated in GPT-4.

In this current work, we incorporate a calligraphy style encoding module into the input part of the generative model to consider the way Chinese characters are written. Then we designed *GlyphGenius*, a Chinese character design

platform. Users are free to redraw a character by its “pictorial meaning”. This enhances engagement and creativity while relieving the burden of system data maintenance. In figure 3 below, we redraw the upper component of the character “京” in the shape of Ancient Chinese city gates and city walls, so that the meaning of “ancient wall” and “capital city” is more intuitive. We aim to help non-native learners perceive the meaning and composition principle of Chinese characters’ constituent parts, thus stimulating their creativity and imagination, leaving a lasting impression.

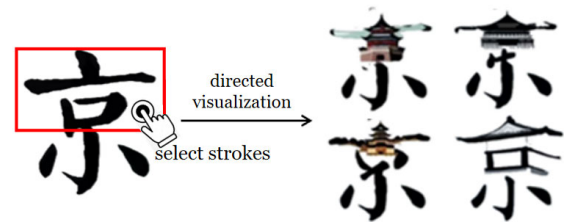


FIGURE 3. The upper parts of the character “京” has been transformed into four different forms under the meaning “ancient city walls”.

The paper’s contributions are as follows:

- Implementation of *GlyphGenius*, an innovative tool with a simplified Graphic-Based UI and a clear operational workflow.
- We present Multi-Stage Model, enabling more precise and superior control over the redrawing of Chinese character components. Additionally, we successfully dissect Chinese characters as distinct components using Scalable Vector Graphics (SVG) files.
- Application of AIGC technology to reconstruct the combination between graphics and Chinese characters, whose pictographic meaning is receding as time goes by, may provide a novel perspective on the learning and cognition of all ideographic symbols.

II. RELATED WORKS

A. TEXT STYLIZATION AND SYNTHESIS

Text stylization and synthesis technologies attempt to generate interesting and clear graphics from given texts. Non-photorealistic rendering method involves changing the topological shape of the text [10], [11], then embedding the text within the graphic outline. Some researchers have concentrated on the challenges associated with the unpredictability that can result from altering characters in their entirety. To mitigate this unpredictability, they sought to dissect and analyze the individual strokes that comprise each character [12]. Other researchers apply variant guided cues [13] or semantic-shape meaning [14] to enhance the controllability of the entire process, which give priority to ensuring the shape of the outline, but to some extent, ignore the readability of the words. Parametric methods represent the style as statistical features, and adjust the target image to satisfy these features [15]. With the emergence of the Generative Adversarial Networks (GAN) [16], style transfer

has gradually become a hot direction, giving rise to numerous studies on Chinese character-based style transfer [17], [18], [19]. This powerful image generation capability of GANs great potential for application in the field of educational visualization. Yet it has seen limited application in Chinese character learning in the past few years.

B. VISUAL GLYPH RENDERING

Researchers have shown the potential of Glyph for visualization in Education [20], [21]. The development of generative modeling now makes this visualization process easier. It is worth noting that there have been considerable endeavors focused on this topic, such as MetaGlyph [22], GlyphControl [23] and GlyphByT5 [24]. However, these methods are limited to handling a whole word or text sequence. In contrast, we aim to empower users to dissect the glyph based on their own idea, particularly when dealing with Chinese characters with multiple strokes and complex structures, setting forth an ambitious goal in Chinese Character visualization field.

C. IMAGE GENERATION BY ARTIFICIAL INTELLIGENCE

The field of image synthesis has advanced significantly in the past few years with the continuous development of Artificial Intelligence Generate Content (AIGC).

Generative Adversarial Networks (GAN) [16] allow for efficient sampling of high-resolution images with good perceptual quality [25], [26]. Experiments have shown that many generative tasks can be effectively accomplished using GANs at low cost, such as text2images [27], [28], image2image [29] and even video generation [30]. Recently, Diffusion Model (DM) [31] has achieved state-of-the-art on most image synthesis tasks. Studies have shown that DMs perform even better than GANs on image synthesis [32]. DM is a class of likelihood-based models which has recently shown potential in a variety of domains, ranging from high-quality text2images [33] and image2image generation [34] to image restoration [35], [36] and image editing tasks [37].

Numerous methods have been proposed to improve DMs. Among these new contributions, Latent Diffusion Model (LDM) [7] works on the latent space of the image instead of the pixel space of the image, which can improve the model's inference speed and reduce the memory consumption. Improvements to LDM over the past two years have largely been driven by improving U-net backbone [38], attention mechanism [34], [39] and Hypernetwork fine-tuning [40]. To ensure robust generalization, our Multi-stage model expands upon the capabilities of the pre-trained Stable Diffusion [9] by adapting it for stroke-based and multi-stage training proposed in this paper.

While these efforts have shown substantial improvements in generation performance, it is disappointing to note that compared to Chinese characters, their performance in dealing with English fonts is significantly more excellent. In this study, we aim to design a platform for shape-based redrawing of modern logogram, especially Chinese characters, to assist in logogram language learning.

III. SYSTEM OVERVIEW

A. MOTIVATION

The inspiration behind our integration of AI and Chinese character design originates from the revolutionary impact of AIGC on human-machine interaction paradigms [25], [41]. Coincidentally, when we trace the cultural roots of Chinese characters, we also find a special relationship between “image” and “context”. The earliest hieroglyphic origins of the Chinese characters are that people used vivid images as words to convey complex natural language messages. Therefore, Chinese characters not only convey a certain meaning, but also look like the meaning it express.

Thus, we put forward the idea to rebuild the relationship between image and the pictorial richness of logograms through generative capabilities of AIGC. As a result, we have identified the target user group, namely non-native Chinese language learners, of our system as an engaging Chinese language learning platform. In line with this goal, we have defined the following philosophy in our design:

1) GRAPHIC-BASED UI

In order to allow such users to interactively deal with the design requirements, we apply graphics-based UI elements, e.g. block [42]. Graphics-based UI elements simplify the programming vocabulary by choosing components from palette [42]. Because we aim to bridge the gap between images and natural language for users, a non-code flow must be implemented.

2) SIMPLIFIED OPERATION

To highlight the important information and help users form short-term memory, unnecessary operational processes must be minimized. Users can invoke the generative model and OCR model by simply touching and inputting information. For example, during the stage where users handwrite Chinese characters, the system can automatically invoke the OCR recognition model and proceed to the next stage if the user's finger has been off the screen for 2 seconds. This interaction method simulates handwriting input methods commonly found on mobile devices, enhancing user comfort.

3) MORE INTERACTIVE

We allow the user to write any Chinese characters and select any strokes that he or she wants to deform for visualisation. Subsequently, the user assigns visual semantics to the selected stroke group that constitute these components. After the image generation is completed, users have the freedom to move and scale it, when combining it with the unselected stroke group, until they reach the most satisfactory state. We also allow users to rewrite, reselect strokes, and choose alternative generated results during the design process. Increased interactivity enhances user engagement, thereby leaving a lasting impression.

B. OVERVIEW

The architecture of our system *GlyphGenius* with layers from top to bottom is depicted in figure 5.

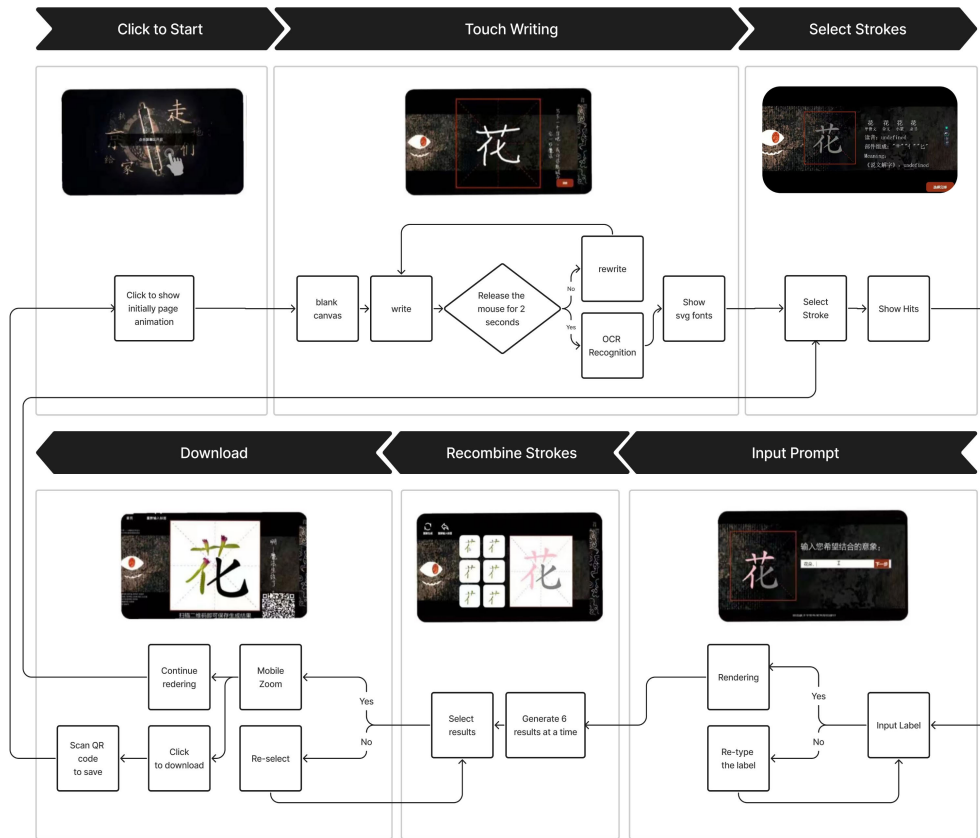


FIGURE 4. The operational workflow is divided into 6 steps: Click to Start, Touch Writing, Select Strokes, Input Prompt, Recombine Strokes and download.

1) APPLICATION LAYER

The system is developed using HTML and is compatible with various Internet-connected interactive devices such as smartphones, computers, and touchscreens. It is deployed on a lightweight ECS. The operational workflow of the system comprises six steps, which will be elaborately explained with illustrations in the next subsection III-C.

2) SERVICE LAYER

Given that user interactions on the website involve invoking models, the service layer is deployed separately on GPU workstations with high computational capabilities. The application layer can send requests through the Axios library to trigger the model services in this layer and obtain results, thereby achieving a clear separation between front-end and back-end. The OCR model [43] is employed to recognize Chinese characters written during the “touch writing” step in the application layer. The multi-stage model is the core component of the system and will be detailed in III-D2. A module that converts images into publicly accessible URLs enables the results of the “recombine strokes” step in the application layer to be transformed into public web links, and subsequently into QR codes using a QR Code generator library. This facilitates users in downloading their design results.

3) DATA LAYER

At the bottom, the Data Layer provides data storage services for the website. The metadata of Chinese characters is stored in JSON format. During the “select strokes” step in the Application layer, it is displayed on the screen to help users understand more about the given Chinese character. The SVG files are a large dataset of Chinese characters generated in advance from TTF files [44], [45], [46]. They are also stored in SVG file format for ease of use in the “select strokes” stage of the interaction layer. The specific application and algorithm will be detailed in III-D1.

C. USAGE SCENARIO

The operation process in the application layer is divided into six steps. Figure 4 shows the detailed workflow. The first three steps are designed for users to determine which character and which part of it to redraw; the last three steps are designed to perform directional deformation and recombination. We take character “花” as a case study below, wherein the workflow is delineated in meticulous detail:

Step1 Wake up the screen by clicking it.

Step2 Write a character. The system supports both touch and mouse click interactions. The center of the screen displays a grid for writing the character that the user wants to design. Upon completion of handwriting Chinese characters,

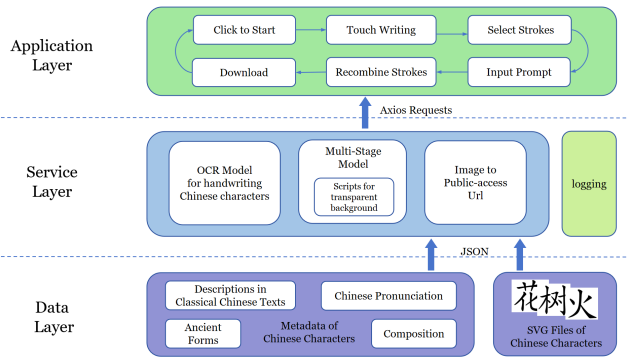


FIGURE 5. The layered architecture of GlyphGenius, detailing the Application Layer, Service Layer and Data Layer from top to bottom.

this system automatically triggers OCR recognition and proceeds to the subsequent stage. an alert window will be displayed to indicate any discrepancies if something goes wrong with recognition.

Step3 Select strokes. On the left column, the system display the SVG file of the character recognized in step 2, allowing users to interactively select specific strokes by clicking on them. Upon selection, the stroke’s color changes, and a subsequent click cancels the selection. Moreover, we provides additional information about this Chinese character in the right column, such as its pronunciation, component composition, English meaning and interpretations from the Chinese classic text *Shuo Wen Jie Zi*. These details aid in associative memory and serve as a reference for providing cues for entering deformation prompt in step 4.

Step4 Input the prompt. During this stage, the selected stroke group in step 3 continuously flicker to indicate the component that will be deformed. Users can enter the prompt based on this character’s meaning provided in Step 3, or their own creative (e.g. “blossoming flowers with green leaves”).

Step5 Choose the most satisfactory redrawn result from the six generated options. Alternatively, users can choose to “rerender” or “reenter labels” if they are not satisfied with the generated results. After a result is chosen, users can freely move and zoom to recombine the chosen result with the unselected stroke group in Step 3 until they achieve the desired effect. Then, click the download button to proceed to the next step.

Step6 Download the artistic wordart via QR code. Upon completion of the download, the system allows users to either return to the homepage at step 1 or to continue the design process for this specific character.

In figure 6, we also performed this visual transformation on other Chinese characters. For instance, by selecting all the strokes of the character “人” and entering the prompt “a woman walking with strides,” we obtained a vivid depiction of a woman in motion.

D. METHOD

1) SELECT BY STROKES

A Chinese character is usually composed of several different constituent parts, and the meanings of these parts are likely to

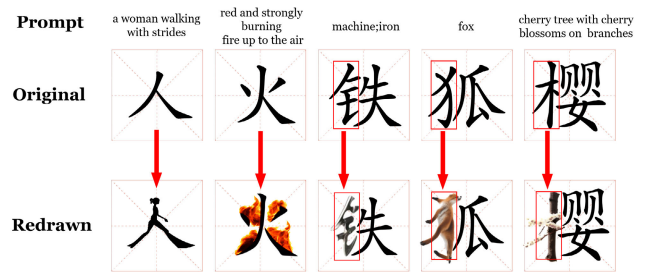


FIGURE 6. Other cases designed by our system, including redrawing the whole character and its constituent parts.

be different. So separating by parts and rendering respectively can avoid confusion of different part meanings. Considering Chinese characters are written based on stroke order, which means that strokes are considered as the smallest unit of writing. Therefore, we set strokes as the minimum chosen unit.

There have been some works using segmentation to create new glyphs or fonts from existing Chinese characters [47] and bitmap images of handwriting [48]. However, their approach is highly dependent on an automatic thinning-based stroke segmentation method that unfortunately does not perform well with fonts.

We observe the use of SVG files, commonly used in icon design and data visualization. Numerous studies have already explored the utilization of parametric methods for stroke extraction models in Chinese characters [44], [45]. Building upon this foundation, we are able to process TrueType Font (TTF) files [46] and generate SVG files of the corresponding Chinese characters. In the SVG file format, each stroke in Chinese character is controlled by closed Bézier curves. Therefore, it’s a simple but efficient method to differentiate strokes of a single Chinese character, using a script to batch-modify the filling colors of the strokes in a SVG file. We assign sequentially decreasing colors to each stroke based on its writing sequences.

Following the implementation of this process, we apply algorithm1 to accomplish selection and deselection on strokes. Consequently, users can select the strokes they want to redraw (click again to deselect), thereby enhancing the flexibility and interactivity of the deformation process. Once the user confirms their selection, the system generates a separate image A composed of the selected strokes, ready to be sent to the model as a base image. Simultaneously, the unselected strokes collectively form another image B, which will later be shown in step 6 of section III-B to combine with the output of the model for image A.

2) MODEL ADAPTATION

Stable Diffusion has been proved very effective in image synthesis. Low-Rank Adaptation (LoRA) [49] prevents catastrophic forgetting by learning the offset of parameters with low rank matrices, based on the observation that many over parameterized models reside in a low intrinsic dimension subspace [40], [50]. It allows optimization of parameters in a

Algorithm 1 Select Strokes

```

1: Input: SVG file with sequentially decreasing RGB color
   values in stroke order.
2: Output: The list of selected stroke index.
3:  $N \leftarrow$  total number of strokes in svgFile
4:  $S \leftarrow$  empty list
5:  $X \leftarrow$  RGB value of original strokes
6:  $Y \leftarrow$  RGB value of selected strokes
7: repeat
8:   if clicked then
9:      $Color \leftarrow getRGBValue(\text{current pixel})$ 
10:    if  $Color \in [X, \dots, X + strokeNum]$  then  $\triangleright$  click
       on an unselected stroke
11:      $Index \leftarrow Color - X$ 
12:     Push Index into S  $\triangleright$  Adjust the display
       effect
13:     Update stroke color in svgFile as  $Y+index$ 
14:    else if  $Color \in [Y, \dots, Y + strokeNum]$  then  $\triangleright$ 
       click on a selected stroke
15:      $Index \leftarrow Color - X$ 
16:     Remove Index from S  $\triangleright$  Adjust the display
       effect
17:     Update stroke color in svgFile as  $X+index$ 
18:    end if
19:  end if
20: until the 'next' button is clicked
21: return S

```

certain subspace while still maintaining the model's learning performance with fewer samples and lower computational power support.

We conducted our experiments based on the capabilities of the pre-trained Stable Diffusion V1.5 [9]. To further control the generation effect, we use LoRA to adjust the parameters of the model. For a start, we trained LoRA to generate some special styles to assist in the design of the GlyphGenius, such as flame, flower etc. As is shown in figure 7 (a), the design results generated by the original model without LoRA, whose color has changed only. Apparently, in figure 7 (b)

and (c), after applying LoRA, the visual transformation is more detailed.

During the experiment, as is shown in figure 7 (b), we found that the structure of many generated characters has changed a lot. We assume that The large-scale model perceives Chinese characters as images rather than individual strokes. Consequently, although the generated results are visually appealing, there is a significant decrease in the recognizability. [51] has demonstrated that employing a staged generation enhances the realism and detail of the generated images. To maintain the rigor and beauty of the redrawn output in the character structure, we use two models with different emphases, and render the calligraphy characters in two stages in chronological order. First, we record the user's writing sequences when the user handwrites on the screen. Every time the finger leaves the touch screen, the system judges that a stroke is finished and captures the handwriting on the screen, thus recording the user's writing sequences. Once the user has entered the prompt in step 4, the system sends the writing sequences with the prompt and the selected stroke group together to the Multi-stage model.

Figure 8 shows the architecture of the proposed Multi-stage model. In the first stage, we use a specific model that focuses more on details. According to the user's writing sequences, we redraw the corresponding strokes in chronological order. In this stage, based on the writing sequence and input prompts from user, the model focuses on understanding Chinese characters in terms of strokes rather than a complete image. This approach ensures attention to the details of strokes and stroke order, while maintaining the stability of the character's structure.

In the second stage, we use another model that focuses more on the font structure to render and redraw the strokes that have been individually rendered in the first stage as a whole. Based on the calligraphy characters designed in the first stage and input prompt from user, the model generates new images, automatically adjust the redraw parameters, generate images that conform to the structure of calligraphy characters to ensure a more consistent style of each stroke and achieve an overall aesthetically pleasing result. Our method can effectively maintain the structure and details of calligraphy characters, as well as enhance the controllability and recognizability of the generated results.

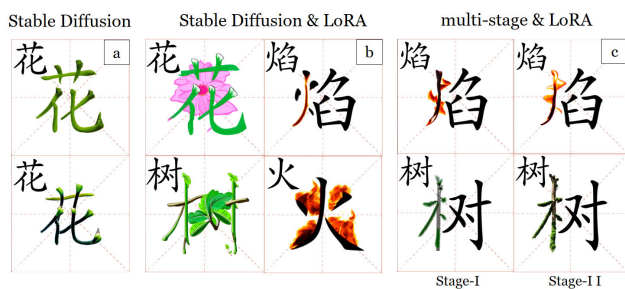


FIGURE 7. (a)A comparison of the design results generated by the original model with and without LoRA. (b)The results of special styles: flame and flower. (c)Two sets of visual comparison graphs. The character “焰” and “树” are generated with the prompt “the blazing flames” and “large tree with green leaves” respectively..

IV. PERFORMANCE EVALUATION

To further examine whether the multi-stage rendering method enhances the controllability and brings Chinese characters closer to real images, we conducted performance evaluation in both qualitative and quantitative aspects. In the qualitative aspect, we conducted ablation study to demonstrate the role of writing sequences and the performance of our multi-stage model visibly. In the quantitative aspect, on one hand, we evaluate the quality of image synthesis by reporting the FID [52] and CLIP-score [53] on CUB [54] and COCO [55] dataset. On the other hand, we assess the controllability of Chinese character redrawing by reporting skeleton similarity index calculated from SSIM [56], MSE [57] and PSNR [58], as well as the recognizability by recognition index [43]. The

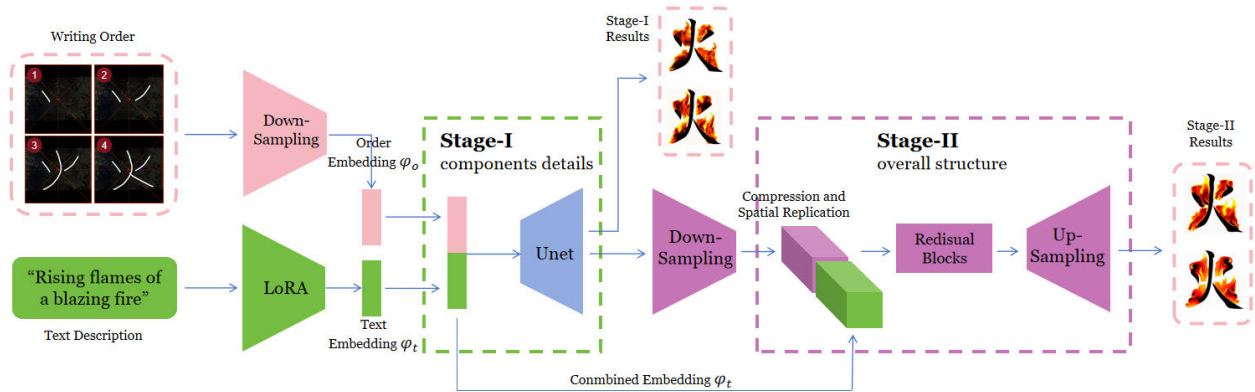


FIGURE 8. The architecture of the proposed Multi-stage model in two stages according to the user's writing sequences, input prompt words, Chinese character structure, and other information.

examination for controllability and recognizability aims to ensure the redrawing outputs of our model do not confuse Chinese character learners.

A. ABLATIVE STUDY

As is shown in figure 9, we present four prompt settings for the case “花” to simulate potential behaviors of users:

- **No Prompt** No specific prompt provided.
- **Insufficient Prompts** Partial prompts that do not fully cover objects in the conditioning images (e.g., the default prompt used in this paper: “a high-quality and detailed masterpiece”).
- **Conflicting Prompts** Prompts that alter the semantics of the conditioning images (e.g., “a delicious cake”).
- **Perfect Prompts** Precise prompts describing necessary content semantics (e.g., “Cherry blossoms”).

We put these four types of prompts into three different models to compare the results: (1) Stable Diffusion V1.5 [9] with LoRA [49]; (2) Stable Diffusion with LoRA adding writing order; (3) Multi-stage generation adding LoRA and writing order. The inclusion of user-specified writing order in the conflicting prompts helps the model maintain better font structures, preserving the information propagation of Chinese characters. Furthermore, after the second-stage rendering, the designed Chinese characters closely align with the provided prompts.

In figure 9, we illustrate the Chinese character “花” with a “top-bottom” structure. Typically, in this type of Chinese character, only one part carries pictographic meaning closely related to the overall meaning of the character, while the other part is associated with the pronunciation of the character. Hence, for visual transformation, only the components with pictographic meaning are selected. In addition, we combine it with the remaining untreated part in the column of “Perfect Prompt”, to compare the visual effects of different models applied to this system's modular redraw. More details and cases about ablative study can be found in the supplementary material.

B. QUANTITATIVE EVALUATION

Evaluating the performance of generative models quantitatively can be challenging, especially when we aim to demonstrate that the multi-stage architecture exhibits enhanced controllability and recognizability in Chinese character generation. In this paper, on one hand, we choose Fréchet Inception Distance (FID) [52] and CLIP-score [53] to evaluate the quality of image synthesis. On the other hand, to further validate the readability of Chinese characters with the multi-stage, we report the similarity of skeleton [56], [57], [58], [59] and OCR recognition accuracy [43].

1) IMAGE SYNTHESIS INDEX

To demonstrate the effectiveness of the proposed method, we compare the result of each stage with the open-source stable diffusion model V1.5 [9].

a: DATASETS

We evaluate our multi-stage model for image synthesis on the COCO [55] and CUB [54] datasets. CUB [54] contains 200 bird species with 11,788 images. To evaluate the generalization capability of our multi-stage model, a more challenging dataset, COCO [55] is also utilized for evaluation. This dataset contains images with multiple objects and various backgrounds. Each image in COCO [55] has 5 descriptions, while 10 descriptions are provided by [60] for every image in CUB [54]. Following the experimental setup in [61], we directly use the training and validation sets provided by COCO [55].

b: EVALUATION METRICS

FID [52] and CLIP-score [53] were recently proposed as metrics that consider not only the synthetic data distribution but also how they compare to the real data distribution. FID [52] directly measures the distance between the synthetic data distribution $p(\cdot)$ and the real data distribution $p_r(\cdot)$. In practice, images are encoded with visual features by the inception model. Assuming the feature embeddings follow a multidimensional Gaussian distribution, the synthetic data's

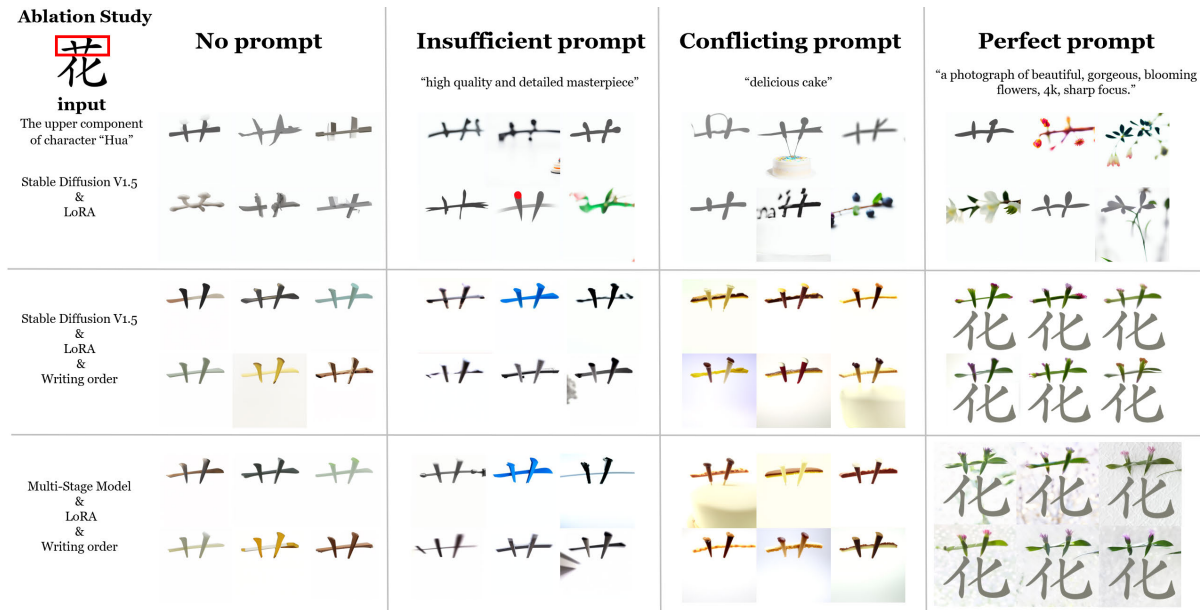


FIGURE 9. The result of the character “花”. This character means flowers. However, its upper component means grass while lower component has irrelevant meaning.

TABLE 1. Evaluation for image generation of two stages compared to Stable Diffusion V1.5 [9] with LoRA [49]. We report FID [52], CLIP-score [53] on both COCO [55] and CUB [54] dataset, average skeleton similarity index metric (*SSIM* [56]) and average recognition accuracy (*OCR – acc*).

Method	FID↓		CLIP-score↑	\overline{SSIM} ↑	$\overline{OCR - acc}$ ↑
	COCO	CUB			
Stable Diffusion V1.5 with LoRA	296.43	282.71	0.261	0.689	0.654
stage-I (Ours)	292.74	283.76	0.235	0.882	0.805
stage-II (Ours)	287.63	280.90	0.258	0.871	0.798

Gaussian with mean and covariance (m, C) is obtained from $p(\cdot)$ and the real data’s Gaussian with mean and covariance (m_r, C_r) is obtained from $p_r(\cdot)$. The difference between the synthetic and real Gaussians is measured by the Fréchet distance, i.e., $FID = \|m - m_r\|_2^2 + Tr(C + C_r - 2(CC_r)^{0.5})$. CLIP-score [53] leverages a pre-trained deep learning model to assess text-image correspondence. The model, called Contrastive Language-Image Pre-training (CLIP), first embeds both the text and image into a common latent space. Each data point is transformed into a high-dimensional vector representing its semantic content. Subsequently, the cosine similarity between these embeddings is computed. The cosine similarity, ranging from -1 to 1 , reflects the alignment between the text and image in the latent space. A CLIP-score [53] close to 1 indicates high semantic similarity, signifying the image effectively portrays the described concept. Lower FID [52] and higher CLIP-score [53] values mean closer distances between synthetic and real data distributions.

c: EXPERIMENT

We first compute the FID [52] score for a unconditional model, 30k 256×256 samples are randomly generated.

To compute the FID [52] score for a text-to-image model, all sentences in the corresponding test set are utilized to generate samples. To better evaluate the proposed methods, especially to see whether the generated images are well conditioned on the given text descriptions, we also conduct user studies. We randomly select 50 text descriptions for each class of CUB sets [54]. For each sentence, 6 images are generated based on the same 256×256 picture of the same character. Subsequently, we utilized these images for CLIP text-image score metrics [53], which signifies the alignment between images and text embeddings in a multi-modal context, reflecting the model’s ability to understand and associate visual and textual information effectively. Notably as shown in the Table 1, both FID [52] and CLIP-scores [53] are better in the second stage. And CLIP-score [53] shows our model well-maintains readability compared to other baselines and also enhances quality compared to the stage-I. As for FID [52], in order to maintain the structural integrity of the original character, the multi-stage model considers user-specified writing sequences and executes a skeleton-based redraw process. This procedure may introduce deviation between the output image and tangible objects. Therefore,





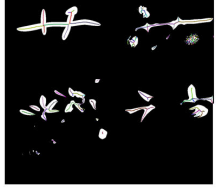



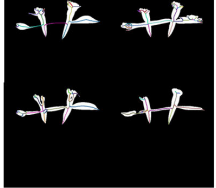
Original	Input	Skeleton	SSIM↑	RSNR↑	MSE↓
 The upper component of character "Hua"			-	-	-
Stable Diffusion V1.5 with LoRA			0.722	14.435	2625.237
Stage-I (Ours)			0.954	17.609	1182.168
Stage-II (Ours)			0.948	16.922	1375.772

FIGURE 10. The process of skeleton extraction and similarity comparison. We take the character “花” as an example. Text descriptions: (1) A blooming Peony. (2) Falling Petals. (3) Blossoming flowers with green leaves. (4) A vast expanse of blooming flowers.

there isn't a significant change in performance compared to the original Stable Diffusion model.

2) SKELETON SIMILARITY INDEX

To better evaluate the proposed methods, especially to see whether the multi-stage model is suitable for Chinese character redraw, we also conduct studies based on the skeleton similarity. The skeleton forms the foundation for accurate recognition of Chinese characters. When characters possess a structured skeleton, they appear more aesthetically pleasing.

a: EVALUATION METRICS

In order to quantitatively assess the skeletal similarity of Chinese characters before and after redraw, we employ a skeleton-tracing algorithm [59] to extract the skeletons of the original and redrawn characters. Subsequently, we measure the similarity between the skeletons using Structural similarity index (SSIM) [56], Peak Signal-to-Noise Ratio (PSNR) [58], and Mean Square Error (MSE) [57] metrics. The most fundamental metric, MSE [57], calculates the average squared difference between corresponding pixels in the original and reconstructed images. Mathematically, for images of size $M \times N$ with original pixels $I(i, j)$ and reconstructed pixels $K(i, j)$: $MSE = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N [I(i, j) - K(i, j)]^2$. Lower MSE [57] indicates better quality, as it reflects a smaller overall discrepancy. PSNR [58] builds

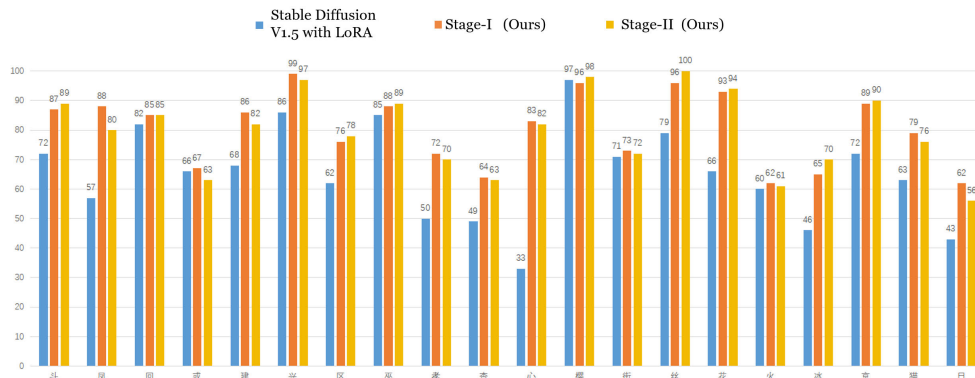
upon MSE [57] by converting it to a decibel scale, a more intuitive measure for signal strength. It is calculated as $PSNR = 10 \log_{10} \left(\frac{max_I^2}{MSE} \right)$, where max_I represents the maximum possible pixel value. A higher PSNR [58] signifies a stronger original signal relative to the introduced noise or distortion. SSIM [56] goes beyond pixel-wise differences, incorporating structural information. It assesses luminance, contrast, and structure similarity between the images. A value closer to 1 indicates greater structural resemblance and higher perceived quality.

b: EXPERIMENT

We selected 10 representative Chinese characters. For each Chinese character, images in 256×256 format with a gray foreground and white background were generated from the font's TTF file [46] and 4 text descriptions were conceived according to its visual and textual semantics. Subsequently, we utilized these text descriptions to guide the model in conditional redrawing based on the generated regular script font images. New images were generated in 256×256 format. Each text description yielded 25 images, amounting to a total of 100 images. To assess the effectiveness of our multi-stage model in managing Chinese character strokes across various stages, we documented the results of (1) stable diffusion V1.5 [9] with LoRA [49], (2) stage-I, and (3) stage-II. Figure 10 illustrates the process of the skeleton extraction and similarity comparison of the character

TABLE 2. Partial Results of evaluation for skeleton similarity of two stages. We report SSIM [56], PSNR [58], MSE [57] for our method of multi-stage generation (stage-I and stage-II) in contrast to original stable diffusion pre-trained model (SD) [9].

	SSIM \uparrow			PSNR \uparrow			MSE \downarrow		
	SD	Stage-I	Stage-II	SD	Stage-I	Stage-II	SD	Stage-I	Stage-II
火	0.786	0.917	0.917	12.730	14.970	14.498	3550.137	2150.103	2322.065
林	0.827	0.842	0.823	11.106	12.125	11.468	5059.688	4171.295	5037.165
丝	0.738	0.867	0.863	11.254	11.895	11.382	5332.129	4887.307	4940.827
心	0.613	0.876	0.893	11.173	11.618	12.377	5000.029	4507.053	3833.836
花	0.722	0.954	0.948	14.435	17.609	16.922	2625.327	1182.168	1375.772

**FIGURE 11.** The recognition rates of 2600 rendered images for the 20 representative Chinese character classes were compared after multi-stage rendering. Notably, after the second stage of rendering, Baidu OCR demonstrated varying degrees of improvement in recognizing the designed Chinese characters.

“花”. Detailed results can refer to the supplementary materials.

Partial results and their text descriptions are shown in Table 2. Considering that MSE [57] values appear significantly large as they directly measure the distance between two point sets in the given two images and PSNR [58] lacks a defined upper limit, making it challenging to interpret the effectiveness of the metrics, we ultimately chose SSIM [56] for the metric of “Skeleton Similarity Index” in Table 1 above. SSIM [56] ranges between 0 and 1, the average SSIM [56] values of the analysis results for these selected characters are presented as the source of data for \overline{SSIM} in Table 1. In terms of the fact that stage-I achieves higher score than stage-II, we speculate that this might be due to Balancing the fundamental structure of Chinese characters with richer aesthetic features poses a significant challenge. Consequently, the results of stage-II are inevitably at a disadvantage compared to the stage-I as they struggle to maintain the basic integrity of the Chinese character skeleton while incorporating more detailed and realistic aesthetic traits.

3) CHARACTER RECOGNIZABILITY INDEX

To validate the readability of Chinese characters with the multi-stage model, we employed a Chinese character recognition model [43] to identify the output from both stage-I and stage-II in our multi-stage model. The accuracy of recognition was used as an indicator to assess whether

the designed Chinese characters retained readability. In the Unicode encoding, Chinese characters are categorized into 16 distinct structural classes. For Chinese language learners, our system excluded three uncommon structural sequences (U+2FFC, U+2FFE, and U+2FFF). From the remaining sequences, we selected 20 representative Chinese characters to evaluate whether our method enhances the readability of the design.

Initially, for each of the 20 representative Chinese characters, we binarized and converted them to 256×256 dimensions with black foreground on a white background. Then we utilized an open-source Optical Character Recognition (OCR) method [43] for identification, obtaining the recognition rate as a quantitative measure of the recognizability of the generated results. In contrast, we applied a similar process to stable diffusion V1.5 [9] with LoRA [49]. For each Chinese character, 100 results are generated by each model involved in the comparison. The number of results correctly identified by the OCR model [43] represents the generation accuracy of that model for that particular character. Figure 11 shows the results of 20 characters generated from stable diffusion and two stages of our model. The average recognition rate of these 20 Chinese characters is taken as the “OCR-acc” metric in Table 1. The results show that multi-stage architecture achieves relatively higher recognizability in Chinese character generation. During the processing within the model, specifically within stage-II, where results from stage-I undergo further refinement, there was not a significant impact on recognizability. Additionally,

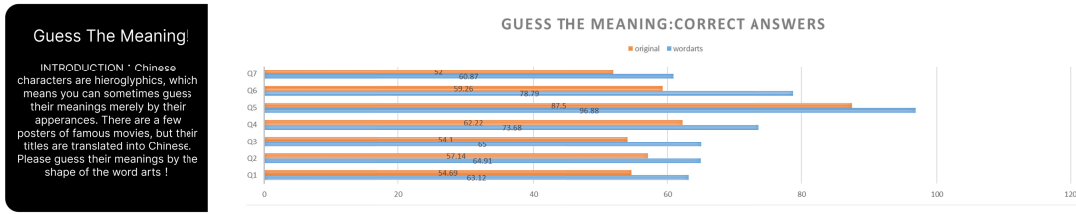


FIGURE 12. The overall result of seven questions in this quiz. Each participant gets either the original character or the “wordarts” randomly in each question.

data such as the CLIP-score [53] indicates that the results from stage-II, in comparison to stage-I, possess greater authenticity. This also indicates the effectiveness of the multi-stage architecture.

V. USER STUDY

A. GUESS-MEANING QUIZ

To quantitatively evaluate the positive impact of Chinese Character learning produced by our approach, we send out online questionnaires.

1) QUIZ DESIGN

In this quiz, we prepare a total of seven questions. Each question corresponds to a Chinese character, and the participants are required to choose one option from the given four options according to their cognition. To minimize the potential impact on participants’ recognition in the control group, we take characters in regular script font for the control group (we refer to as “original group”). On the other hand, we take the redrawn results of the system as the experimental group (we refer to as “wordarts group”). Therefore, we can ensure that difference in recognizing a character in the wordarts group is solely due to its visual form. For instance, in figure 13, the left component is redrawn according to the character’s meaning “wolf” and combined with its original right component, while the original whole character is taken as the control group. Four options are set in the question: Strong, Wolf (correct), Man and Iron. Figure 12 shows the overall accuracy of every question in two groups.

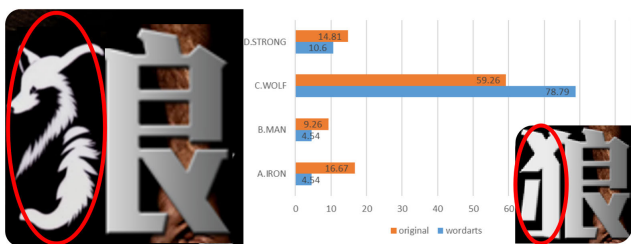


FIGURE 13. The character “狼” in the quiz and its accuracy in both groups.

To avoid one participant seeing the same character in two groups, we randomly assigned each question in the quiz to either the wordarts group or the original group. Namely, each participant only gets either the original character or the “wordarts” in their quiz. We send online questionnaires to international students and collected the

results to quantitatively evaluate the effectiveness of the system on Chinese character learning.

2) FEEDBACK

A total of 135 valid responses are received. Among these participants, 89.55% are new to Chinese language learning(learning within 1 year or never). We refer to this group as “beginner Chinese learners”, while others are referred to “experienced Chinese learners”. Considering that experienced Chinese learners have relatively strong grasp of Chinese characters, we exclude data from them and keep 121 samples of beginner Chinese learners for further analysis.

The scoring results of the statistical answer sheets are presented in figure 14. The accuracy of understanding Chinese character meanings after seeing the redrawn characters is consistently above 60%, with an average increase of 12.76% compared to that after seeing the original characters. The correct rate of the “wordarts” group and the correct rate of the original group. is significantly higher than that of people with Chinese learning experience, i.e., wordarts is able to improve the recognition accuracy of Chinese characters for beginners more significantly than that of people with Chinese learning experience, which proves the potential value of our work in the initial learning of Chinese characters. More details and cases can refer to the supplementary material.



FIGURE 14. The accuracy of these 7 questions among the beginner Chinese learner group (a) and experienced Chinese learner group (b).

B. SYSTEM USABILITY TEST

We set a physical stand with touchscreen in the university and recruited 21 volunteers(11 males and 10 females) for the test. There are 9 international students studying in China among volunteers, 3 of them have learned Chinese longer than one year. Volunteers major in business, architecture, automation, etc. and come from different grade levels. 15 (66.7%) of them have no experience in designing.

1) PROCESS

We divided the workshop into three sessions. We first introduced the purpose of the project we were working on and demonstrated the usage of the system with specific examples. After grasping the workflow under our guidance, volunteers were free to generate characters of their own choice independently. In this session, we also prepared a set of Chinese characters along with their meanings as prompts. For example, the character “花” refers to a flower, with the upper part of the character indicating “草”, the character “火” represents a flame, etc.. Participants were also encouraged to contribute their own ideas and creativity by inputting new Chinese characters or attempting new prompt words. Finally, after a 20-minute free trial, we distributed a survey to all users and collected feedback from the participants on their experience with this system. The questionnaire consisted of seven questions in total:

- **Q1 [workflow]** Was the whole workflow of ‘Draw, Render, Reorganize’ obvious and natural to you?
- **Q2 [performance]** How satisfied you are with the output of this system?
- **Q3 [design]** Do you think you can design creative word arts with your own ideas?
- **Q4 [cognition]** Do you think this system is helpful for the understanding of the original meaning of Chinese character components? Was the whole workflow of ‘Draw, Render, Reorganize’ obvious and natural to you?
- **Q5 [non-expert use]** Do you think this system is friendly to those with little experience in graphic design?
- **Q6 [pedagogy]** Do you think this system helps junior students and international learners in Chinese and Chinese character learning?
- **Q7 [overall]** We expect your overall rating of our system!

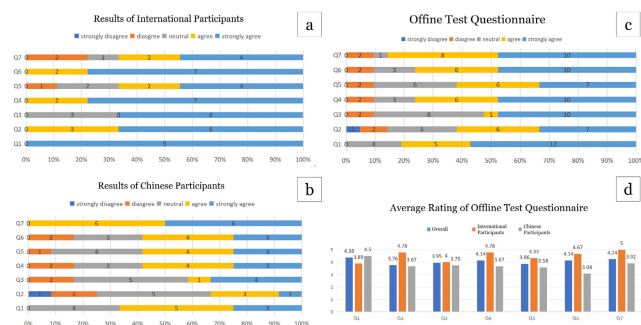


FIGURE 15. The distribution of scores assigned by international (a), Chinese participants (b) and all participants (c) across the seven questions. (d) The average score of each question for overall participants, international participants and Chinese participants.

The answers were rated on a 5-point Likert scale (5.strongly agree - 3.neutral - 1.strongly disagree), which are summarized in figure 15 (d). We analyze the results by dividing them into two groups: the international participant group and the Chinese participant group. Figure 15 (a), (b) and (c) display the average total scores of international participants, Chinese participants and all participants respectively across Q1-Q7 mentioned above.

2) FEEDBACK

Through the offline test, we gained more insights into both advantages and drawbacks of our approach. More details and cases about ablative study can refer to the supplementary material.

For Q1 [Workflow], most participants were satisfied with the system’s workflow, with an average rating of 4.38 out of 5.

For Q2 [Performance], due to the diversity of Chinese characters and stroke structures, the rendering results of the model are significantly varied depending on the character shape, which leads to significant differences in the participants’ evaluations of the model’s performance.

For Q3 [design], almost half volunteers are convinced that the system can assist them in designing creative wordarts.

For Q4 [cognition], Q5 [non-expert use], Q6 [pedagogy], participants generally believe that the system is easy as well as effective for non-native Chinese learners to learn Chinese character components’ meaning and composition principles. Participant 4 from China commented: “Selecting strokes is truly a innovation. For example, the left half of the character of cat contains the meaning of an animal, while the right half does not. This kind of processing is accurate.”

Some participants also raised other shortcomings during use. For example, participant 20 from India noted: “Both the idea and interaction of the system are interesting, but loading and image generation could be faster.”

Overall, most participants were satisfied with the overall system, with an average score of 4.24. We also calculate the result of international participants and Chinese participants respectively. The rating distribution is quite different between Chinese and international participants in figure 15 (a) and (b). Since Our target users are beginner Chinese learners, we speculate that the higher ratings in the international participants may be attributed to their similarity with the target user group, while Chinese participants, with a sufficient understanding of Chinese characters, have a alternatively lower rating.

VI. DISCUSSION

Through the evaluation experiment, we have demonstrated the superiority of our multi-stage model over the Stable Diffusion model [9] in Chinese character generation on both controllability and recognizability, while from the perspective of image synthesis performance, there is no significant difference. Furthermore, we substantiate the effectiveness of our approach and system through the Guess-Meaning experiment and system usability testing. For individuals studying Chinese for less than a year, providing redrawn characters led to an average 12.76% increase in their meaning recognition accuracy compared to the original character, thus enhancing efficiency in learning Chinese characters. In terms of system satisfaction, participants in our workshop rated an average of 4.24, with international participants averaging a score of 5.

One concern about the findings is that our current research focuses on learners with less experience in Chinese. More experienced learners may have developed their own learning strategies, with a more nuanced understanding. Future

research should be undertaken to explore the adaptability and effectiveness of AIGC in supporting their learning needs. Furthermore, relying solely on pictorial meanings for learning Chinese characters is insufficient. Chinese character learning also emphasizes repetitive writing to strengthen memory and comprehension. In future iterations, we intend to integrate a writing practice component into the system. Another major limitation is that the current system still relies on human understanding of Chinese characters for redrawn prompts. And the accuracy and richness of prompts significantly impact the quality of the generated results. For instance, in figure 16 (a), a more detailed prompt can yield better results. Moreover, it is necessary to acknowledge that there are also many Chinese characters that lack pictorial meanings and are unrelated to intention, such as certain predicates and verbs, posing challenges for AIGC applications. To address this limitation, future advancements could involve leveraging large language models to autonomously generate designs based on contextual semantics [10].

In terms of the ability of AIGC to support the learning of characters from languages similar to Chinese, the core principles and techniques employed in our model can also be adapted to other logograms. Unlike AI-assisted learning such as portable devices [5], which solely collect and analyze data, the unique feature of AIGC lies in its ability to generate new learning data. Moreover, by grasping the intrinsic characteristic of pictorial meaning in logograms, we can view characters as visual representations like images, and then utilize AIGC to amplify their interrelations and bring them closer together. We have also conducted research on some other logograms. For instance, Japanese includes hiragana, katakana, and kanji characters, among which Kanji carries the most pictographic meanings. Korean utilizes the Hangul script, which shares visual similarities with Chinese characters. Ancient Chinese scripts like oracle bone scripts inherently possess rich pictorial meanings. We also feed images of these logograms into our multi-stage model, with the utilized prompts and outputs depicted in figure 16 (b)-(d). For their specific application, for instance, integrating handwritten text recognition for them and adding explanations of these logograms to the metadata in the data layer would facilitate a multi-lingual interface.

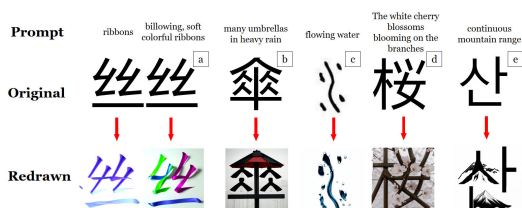


FIGURE 16. Images of logograms and their prompts fed into our multi-stage model: (a) Results generated using the same character (meaning “ribbons”) but different prompts; (b) Traditional Chinese, meaning “umbrella”; (c) Oracle Bone Script, meaning “water”; (d) Japanese Kanji, meaning “cherry blossom”; (e) Korean, meaning “mountain”.

VII. CONCLUSION

In this paper, we present *GlyphGenius*, a pioneering platform that leverages AIGC to assist non-native Chinese learners

in Chinese character learning. The system employs a Multi-stage model, with stroke writing sequence as one of the inputs, to enhance the controllability and recognizability of the generated results. In addition, we propose an interactive and decomposable Chinese character visualization method to help users perceive the principles of character composition and formation. Performance evaluation and user study demonstrated that our approach and interaction are effective and engaging in both quantitative and qualitative aspects. The web page is now accessible via the link <https://scroll.ihanzi.net>.

We are convinced that it’s promising to apply AIGC to connect the relationship between the semantic richness of ideographs and the generative capabilities of AIGC. This work marks a significant stride in the study and cognition of Chinese characters. Future work should encompass the extension of this methodology to other ideographic symbols.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and also would like to thank their sincere gratitude to the International School of Tongji University for administering the online guess-meaning questionnaire.

(Linhui Wang and Yi Lou contributed equally to this work.)

REFERENCES

- [1] J. Zhang, M. Gao, W. Holmes, M. Mavrikis, and N. Ma, “Interaction patterns in exploratory learning environments for mathematics: A sequential analysis of feedback and external representations in Chinese schools,” *Interact. Learn. Environ.*, vol. 29, no. 7, pp. 1211–1228, Oct. 2021, doi: [10.1080/10494820.2019.1620290](https://doi.org/10.1080/10494820.2019.1620290).
- [2] H. Wen, Z. Wang, and Q. Lu, “Extracting Chinese domain-specific open entity and relation by using learning patterns,” in *Proc. ACM Turing Celebration Conf. China*, vol. 17. New York, NY, USA: Association for Computing Machinery, May 2020, pp. 119–125, doi: [10.1145/3393527.3393548](https://doi.org/10.1145/3393527.3393548).
- [3] Y. Yang, L. Zhou, R. Li, H. Yao, J. Song, and F. Ying, “Chinese character learning system,” in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, vol. 21. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–5, doi: [10.1145/3290607.3312813](https://doi.org/10.1145/3290607.3312813).
- [4] Y. Yang, H. Leung, H. P. H. Shum, J. Li, L. Zeng, N. Aslam, and Z. Pan, “CCESK: A Chinese character educational system based on Kinect,” *IEEE Trans. Learn. Technol.*, vol. 11, no. 3, pp. 342–347, Jul. 2018, doi: [10.1109/TLT.2017.2723888](https://doi.org/10.1109/TLT.2017.2723888).
- [5] K. Otsu and T. Izumi, “Interactive handwriting device for enhancing active recollection of character forms by voice assistance for Chinese character learning,” in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, vol. 29. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–7, doi: [10.1145/3544549.3585814](https://doi.org/10.1145/3544549.3585814).
- [6] Y. Ito, T. Terada, and M. Tsukamoto, “A system for memorizing Chinese characters using a song based on strokes and structures of the character,” in *Proc. 17th Int. Conf. Inf. Integr. Web-Based Appl. Services*. New York, NY, USA: Association for Computing Machinery, Dec. 2015, pp. 1–9, doi: [10.1145/2837185.2837235](https://doi.org/10.1145/2837185.2837235).
- [7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685, doi: [10.1109/CVPR52688.2022.01042](https://doi.org/10.1109/CVPR52688.2022.01042).
- [8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with CLIP latents,” 2022, *arXiv:2204.06125*.
- [9] Stability-AI. (2022). *StableDiffusion*. [Online]. Available: <https://github.com/Stability-AI/StableDiffusion>

- [10] S. Iluz, Y. Vinker, A. Hertz, D. Berio, D. Cohen-Or, and A. Shamir, "Word-as-image for semantic typography," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–11, Aug. 2023, doi: [10.1145/3592123](https://doi.org/10.1145/3592123).
- [11] C. Zou, J. Cao, W. Ranaweera, I. Alhashim, P. Tan, A. Sheffer, and H. Zhang, "Legible compact calligrams," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016, doi: [10.1145/2897824.2925887](https://doi.org/10.1145/2897824.2925887).
- [12] D. Berio, F. F. Leymarie, P. Asente, and J. Echevarria, "StrokeStyles: Stroke-based segmentation and stylization of fonts," *ACM Trans. Graph.*, vol. 41, no. 3, pp. 1–21, Jun. 2022, doi: [10.1145/3505246](https://doi.org/10.1145/3505246).
- [13] S. Yang, J. Liu, W. Yang, and Z. Guo, "Context-aware unsupervised text stylization," in *Proc. 26th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 1688–1696, doi: [10.1145/3240508.3240580](https://doi.org/10.1145/3240508.3240580).
- [14] J. Zhang, Y. Wang, W. Xiao, and Z. Luo, "Synthesizing ornamental typefaces," *Comput. Graph. Forum*, vol. 36, no. 1, pp. 64–75, Jan. 2017, doi: [10.1111/cgf.12785](https://doi.org/10.1111/cgf.12785).
- [15] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2770–2779. [Online]. Available: <https://ieeexplore.ieee.org/document/8099779>
- [16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680. [Online]. Available: <https://dl.acm.org/doi/10.5555/2969033.2969125>
- [17] Y. Li, G. Lin, M. He, D. Yuan, and K. Liao, "Layer similarity guiding few-shot Chinese style transfer," *Vis. Comput.*, vol. 40, no. 4, pp. 2265–2278, Jun. 2023, doi: [10.1007/s00371-023-02915-w](https://doi.org/10.1007/s00371-023-02915-w).
- [18] Q. Wen, S. Li, B. Han, and Y. Yuan, "ZiGAN: Fine-grained Chinese calligraphy font generation via a few-shot style transfer approach," in *Proc. 29th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 621–629, doi: [10.1145/3474085.3475225](https://doi.org/10.1145/3474085.3475225).
- [19] X. Zhang, Y. Li, Z. Zhang, K. Konno, and S. Hu, "Intelligent Chinese calligraphy beautification from handwritten characters for robotic writing," *Vis. Comput.*, vol. 35, nos. 6–8, pp. 1193–1205, Jun. 2019, doi: [10.1007/s00371-019-01675-w](https://doi.org/10.1007/s00371-019-01675-w).
- [20] M. Boucher, B. Bach, C. Stoiber, Z. Wang, and W. Aigner, "Educational data comics: What can comics do for education in visualization?" in *Proc. IEEE VIS Workshop Vis. Educ., Literacy, Activities (EduVis)*, Oct. 2023, pp. 34–40, doi: [10.1109/eduvis60792.2023.00012](https://doi.org/10.1109/eduvis60792.2023.00012).
- [21] S. Suh, C. Latulipe, K. J. Lee, B. Cheng, and E. Law, "Using comics to introduce and reinforce programming concepts in CS1," in *Proc. 52nd ACM Tech. Symp. Comput. Sci. Educ.* New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 369–375, doi: [10.1145/3408877.3432465](https://doi.org/10.1145/3408877.3432465).
- [22] L. Ying, X. Shu, D. Deng, Y. Yang, T. Tang, L. Yu, and Y. Wu, "MetaGlyph: Automatic generation of metaphoric glyph-based visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 1, pp. 331–341, Jan. 2023, doi: [10.1109/TVCG.2022.3209447](https://doi.org/10.1109/TVCG.2022.3209447).
- [23] Y. Yang, D. Gui, Y. Yuan, W. Liang, H. Ding, H. Hu, and K. Chen, "GlyphControl: Glyph conditional control for visual text generation," in *Advances in Neural Information Processing Systems*, vol. 36, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Red Hook, NY, USA: Curran Associates, 2023, pp. 44050–44066. [Online]. Available: <https://dl.acm.org/doi/10.5555/3666122.3668034>
- [24] Z. Liu, W. Liang, Z. Liang, C. Luo, J. Li, G. Huang, and Y. Yuan, "Glyph-ByT5: A customized text encoder for accurate visual text rendering," 2024, *arXiv:2403.09622*.
- [25] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.
- [26] J. Zhu, L. Gao, J. Song, Y.-F. Li, F. Zheng, X. Li, and H. T. Shen, "Label-guided generative adversarial network for realistic image synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3311–3328, Mar. 2023, doi: [10.1109/TPAMI.2022.3186752](https://doi.org/10.1109/TPAMI.2022.3186752).
- [27] H. Ku and M. Lee, "TextControlGAN: Text-to-image synthesis with controllable generative adversarial networks," *Appl. Sci.*, vol. 13, no. 8, p. 5098, Apr. 2023, doi: [10.3390/app13085098](https://doi.org/10.3390/app13085098).
- [28] S. Xiao, S. Huang, Y. Lin, Y. Ye, and W. Zeng, "Let the chart spark: Embedding semantic context into chart with text-to-image generative model," *IEEE Trans. Vis. Comput. Graphics*, vol. 30, no. 1, pp. 284–294, Jan. 2023, doi: [10.1109/TVCG.2023.3326913](https://doi.org/10.1109/TVCG.2023.3326913).
- [29] S. Xie, Y. Xu, M. Gong, and K. Zhang, "Unpaired image-to-image translation with shortest path regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10177–10187, doi: [10.1109/cvpr52729.2023.00981](https://doi.org/10.1109/cvpr52729.2023.00981).
- [30] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18444–18455, doi: [10.1109/cvpr52729.2023.01769](https://doi.org/10.1109/cvpr52729.2023.01769).
- [31] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfc8584af0d9671ab10179ca4b-Paper.pdf
- [32] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., Red Hook, NY, USA: Curran Associates, 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf
- [33] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:245335086>
- [34] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7667–7677, doi: [10.1109/iccv51070.2023.00708](https://doi.org/10.1109/iccv51070.2023.00708).
- [35] M. Ren, M. Delbracio, H. Talebi, G. Gerig, and P. Milanfar, "Multiscale structure guided diffusion for image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10687–10699, doi: [10.1109/iccv51070.2023.00984](https://doi.org/10.1109/iccv51070.2023.00984).
- [36] Y. Miao, L. Zhang, L. Zhang, and D. Tao, "DDS2M: -supervised denoising diffusion spatio-spectral model for hyperspectral image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12052–12062, doi: [10.1109/iccv51070.2023.01110](https://doi.org/10.1109/iccv51070.2023.01110).
- [37] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DiffEdit: Diffusion-based semantic image editing with mask guidance," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023. [Online]. Available: <https://openreview.net/forum?id=31ge0p5o-M->
- [38] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4172–4182, doi: [10.1109/ICCV51070.2023.00387](https://doi.org/10.1109/ICCV51070.2023.00387).
- [39] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.
- [40] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," 2020, *arXiv:2012.13255*.
- [41] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, Apr. 2024, doi: [10.1145/3626235](https://doi.org/10.1145/3626235).
- [42] D. Bau, J. Gray, C. Kelleher, J. Sheldon, and F. Turbak, "Learnable programming: Blocks and beyond," *Commun. ACM*, vol. 60, no. 6, pp. 72–80, May 2017, doi: [10.1145/3015455](https://doi.org/10.1145/3015455).
- [43] Z.-R. Wang, J. Du, and J.-M. Wang, "Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107102, doi: [10.1016/j.patcog.2019.107102](https://doi.org/10.1016/j.patcog.2019.107102).
- [44] M.-W. Li, Y. Yu, Y. Yang, G. Ren, and J. Wang, "Stroke extraction of Chinese character based on deep structure deformable image registration," in *Proc. Conf. Artif. Intell. (AAAI)*, 2023, pp. 1360–1367, doi: [10.1609/aaai.v37i1.25220](https://doi.org/10.1609/aaai.v37i1.25220).
- [45] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, Nov. 2019, Art. no. 101723. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0933365719302295>
- [46] H. Y. F. Library. *Hanyikaiti*. [Online]. Available: <https://www.hanyi.com.cn/productdetail?id=814&type=0>

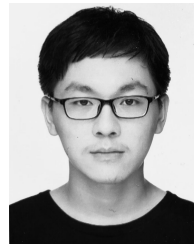
- [47] W. Wang, Z. Lian, Y. Tang, and J. Xiao, "DeepStroke: Understanding glyph structure with semantic segmentation and Tabu search," in *MultiMedia Modeling*, Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, Eds., Springer, 2020, pp. 353–364. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-37731-1_29
- [48] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015. [Online]. Available: <https://www.science.org/doi/abs/10.1126/science.aab3050>
- [49] 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," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [50] C. Li, H. Farkhoor, R. Liu, and J. Yosinski, "Measuring the intrinsic dimension of objective landscapes," 2018, *arXiv:1804.08838*.
- [51] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019, doi: [10.1109/TPAMI.2018.2856256](https://doi.org/10.1109/TPAMI.2018.2856256).
- [52] 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," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6629–6640. [Online]. Available: <https://dl.acm.org/doi/10.5555/3295222.3295408>
- [53] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIP-Score: A reference-free evaluation metric for image captioning," 2021, *arXiv:2104.08718*.
- [54] C. Wah, S. Branson, P. Welinder, P. Perona, and S. J. Belongie. (2011). *The Caltech-UCSD Birds-200–2011 Dataset*. [Online]. Available: <https://api.semanticscholar.org/CorpusID:16119123>
- [55] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755. [Online]. Available: <https://api.semanticscholar.org/CorpusID:14113767>
- [56] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).
- [57] J. Ren, M. Zhang, C. Yu, and Z. Liu, "Balanced MSE for imbalanced visual regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7916–7925, doi: [10.1109/CVPR52688.2022.00777](https://doi.org/10.1109/CVPR52688.2022.00777).
- [58] F. A. Fardo, V. H. Conforto, F. C. de Oliveira, and P. S. Rodrigues, "A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms," 2016, *arXiv:1605.07116*.
- [59] LingDong. (2020). *Skeleton-Tracing*. [Online]. Available: <https://github.com/LingDong/-skeleton-tracing>
- [60] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58, doi: [10.1109/CVPR.2016.13](https://doi.org/10.1109/CVPR.2016.13).
- [61] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 1060–1069. [Online]. Available: <https://dl.acm.org/doi/10.5555/3045390.3045503>



YI LOU is currently pursuing the bachelor's degree in automation with Tongji University. His research interests include computer vision, machine learning, and large-scale AI models.



XINYUE LI is currently pursuing the bachelor's degree with Tongji University. Her research interests include machine learning, image processing, and the application of artificial intelligence in Chinese culture.



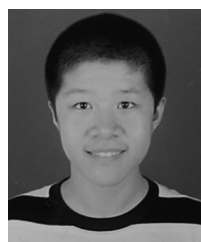
YUXUAN XIANG is currently pursuing the bachelor's degree in environmental design with the College of Design and Innovation, Tongji University. His research interests include user experience research, human-computer interaction research, and combining newest artificial technology with Chinese culture, including machine learning and image processing.



TIANYI JIANG is currently pursuing the bachelor's degree in digital media design with Tongji University. She has extensive experience in the field of UI design and technology development and is passionate about innovation and exploration.



YIYING CHE is currently pursuing the bachelor's degree in computer science and technology with Tongji University. Her research interests include computer vision and pattern recognition.



LINHUI WANG is currently pursuing the bachelor's degree in computer science and technology with Tongji University. His research interests include style transfer and computer graphics.



CHEN YE (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in computer science from Tongji University. He is currently a Professor with the College of Electronic and Information Engineering, Tongji University. His research interests include machine learning, image processing, big data analysis, and its application in the field of industrial intelligence.

• • •