A Multimodal Language Learning System for Chinese Character Using Foundation Model

Jinglei Yu
School of Educational Technology, Beijing Normal University, China
yujinglei@mail.bnu.edu.cn

Zitao Liu
Guangdong Institute of Smart Education, Jinan University, China
liuzitao@jnu.edu.cn

Mi Tian
TAL Education Group, China
tianmi@tal.com

Deliang Wang
Faculty of Education, The University of Hong Kong, Hong Kong
wdeliang@connect.hku.hk

Yu Lu
Advanced Innovation Center for Future Education, Faculty of Education, Beijing Normal University, China
luyu@bnu.edu.cn

ABSTRACT

Learning Chinese character with multiple definitions is challenging for beginners, while images could help learners get quick understanding and strengthen the memory. To solve the problem, we design a multimodal language learning system for Chinese character featured with AI-generated image definitions. The images with desired semantic meanings are generated by text-to-image foundation model ERNIE-ViLG 2.0. To improve learners' understandings of Chinese character definitions, the system could serve as a knowledge building environment. Learners are expected to contribute ideas collaboratively by voting for the appropriate AI-generated image definitions and choosing to add new qualified ones. The system has been implemented on a mobile application, and future works about estimating and optimizing the built system are discussed.

Keywords

Text-to-image generation, Language learning, Knowledge building

1. INTRODUCTION

Through three thousand years of evolvement, each Chinese character tends to have multiple definitions with original and derived meanings, which is challenging for non-native speakers or even young native speakers to understand and remember. The language and linguistics study found that images could be used as non-verbal mediators, which helps learners build efficient connections between the information and the concepts in memory [8]. In addition, psychologist Allan Paivio proposed dual coding theory [9], which indicates the equal importance of verbal and visual information processing for human, and finds out that visual information could contribute to better memory. Particularly, it has been proved that learners could remember definitions of words better when exposed to both visual and verbal information in second-language learning [10].

We thus propose a multimodal language learning system for Chinese character with both text and image definitions. While the images can be retrieved online, it is hard to guarantee the proper images with the desired meanings could be acquired from the massive online resources. To tackle the issue, we utilize text-to-image generation method to provide the desired images directly from text definitions. Text-to-image generation is one type of AI generation methods, and the cutting-edge enabling technology is based on foundation model (or called pre-trained model) [1]. The capability of foundation model covers language, vision, speech and reasoning, etc. Generally, foundation models are pre-trained on large-scale data and could be flexibly adapted to different downstream tasks via transfer learning, so as to achieve excellent performance. Especially, zero-shot transfer is a feasible way to adapt the model to downstream tasks without tuning parameters. With the help of prompt engineering [7], the foundation model could be fixed and the prompts are used to trigger the model. Conditioned on well-designed text prompt, the text-to-image generation foundation models could create desired and original images. In addition to text-to-image generation, foundation models are capable on many other tasks, such as text-to-text generation (e.g., GPT-3 [2]), image-to-text generation (also known as image caption, e.g., BLIP-2 [6]), text-image pairing (e.g., CLIP [11]), text-to-video generation (e.g., CogVideo [5]), etc.

By leveraging on the foundation model, the system could encourage learners to develop ideas towards the generated images. Specifically, for the same text input, the text-to-image generation foundation model could randomly generate various images. The system supports learners to vote for the appropriate images from all the generated ones. The image with the most votes would be shown at the top of the list for the following learners. In addition, learners could choose to

J. Yu, Z. Liu, M. Tian, D. Wang, and Y. Lu. A multimodal language learning system for chinese character using foundation model. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 520–524, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115764

generate new images and decide whether to add them to the image list as candidates. Based on learners' collaboration, the system could serve as a knowledge building environment to help build community knowledge of Chinese character's definitions.

2. SYSTEM DESIGN

2.1 System Framework and Workflow

The system is a multimodal language learning system for Chinese character featured with AI-generated image definitions. As shown in Figure 1, learners could *input* query Chinese character through user interface, and the system would search Xinhua dictionary's online library via requesting API. Xinhua dictionary, also known as modern Chinese character dictionary, is one of the most authoritative reference books in China. The response is basic text information of the character, including pinyin as phonetic symbol, radical partially indicating semantic meaning, structure representing the stroke composition method, and its multiple definitions. Each definition contains the description of the meaning and its sample words.

Through the user interface, the designed system provides several intelligent functionalities for learners. Firstly, learners could *choose* each text definition to show its image definition. The image definition is directly generated from the text definition by means of foundation model ERNIE-ViLG 2.0. To be specific, ERNIE-ViLG 2.0 [3] is a knowledge-enhanced large-scale Chinese text-to-image diffusion model with 24B parameters, developed by Baidu Inc., China. The diffusion model [4] contains forward and reverse diffusion processes. In forward process, the model gradually adds noises to the image data. While in the reverse process, the model is trained to learn how to denoise and reverse the process to generate the desired image. Based on the basic diffusion model, ERNIE-ViLG 2.0 integrates textual and visual knowledge into the training process to help model focus on important elements, such as critical semantics of texts and salient regions of images. In addition, ERNIE-ViLG 2.0 proposes the Mixture-of-Denoising-Experts (MoDE), which contains multiple "experts" adjusting characteristics of different denoising steps in reverse diffusion process. The performance of ERNIE-ViLG 2.0 is state-of-the-art on text-to-image generation task of zero-shot FID-30K from MS-COCO dataset.

Secondly, learners could also choose to *generate more* images via ERNIE-ViLG 2.0 for the chosen text definition. The newly generated image would show up to the user interface and ask learners to *judge* whether it is appropriate enough to add to the image list. All generated images would be saved to the database as backup, while only the learners confirmed ones could be displayed on the user interface.

Thirdly, learners are encouraged to develop ideas towards the image definitions by voting *like* for the most suitable one. The number of likes would be counted and saved as a key feature of the generated image in the cloud database. The image definitions of the chosen text definition would be displayed on the user interface ranked by the number of likes.

2.2 Text-to-Image Generation Performance

We demonstrate the text-to-image generation performance with an example of Chinese character "Yuan" that has three definitions. As mentioned before, each definition is combined by the description of the meaning and its sample words. The translations of the three definitions of character "Yuan" are shown as the followings:

Definition 1. **Description**: A place where fruits, vegetables, flowers and trees are grown. **Sample words**: Garden. Gardener. Gardening. Garden beds.

Definition 2. **Description**: Originally, it refers to the villa and resting place, and now it refers to the public place for people to play around and entertain. **Sample words**: The Old Summer Palace. Park.

Definition 3. **Description**: Originally, it refers to the tombs of emperors, princes, concubines and princesses of the past generations. **Sample words**: Temple Garden (the ancestral temple built in the graveyard of the emperor). Mausoleum (the tomb of the emperor).

Since the text-to-image generation requires well pre-trained foundation model and efficient computing resource for model inference, we take advantage of Baidu ERNIE-ViLG 2.0 API, and build local server to pre-process text prompt and request the API. The construction of the text prompt is important to the text-to-image generation model, which generally requires two main parts, namely painting object and painting style.

For the painting object part, we investigate two categories of text prompts with help of characters' definitions, which are *description only* and *both description and sample words*. Taking the **Definition 2** of character "Yuan" as an example, we request the two categories of text prompt separately. The results are shown in Figure 2, where Figure 2(b) shows more proper results with *both description and sample words* as text prompt. To be specific, the presentation of Figure 2(a) focuses on the non-critical word "villa" from the description, while Figure 2(b) gets well understanding of both "park" from sample words and "public place for people to play around and entertain" from the description. It may because the description tends to be abstract, while the sample words could provide more specific hints.

For the painting style part, in addition to the realistic style utilized in Figure 2, we also explore various artistic styles like surrealism, conceptual art, impressionism, and different production styles like computer graphic style, illustrator style and pixel style, as shown in Figure 3. Considering about the generalization issues for various Chinese character, we set realistic style for all the image generation, but the system designers or even learners could also make their own choices if needed.

Finally, we identify both description and sample words for painting object part and realistic style for painting style part to construct the text prompt. Four exemplary generated images corresponding to **Definition 1** and **Definition 3** of character "Yuan" are shown in Figure 4.

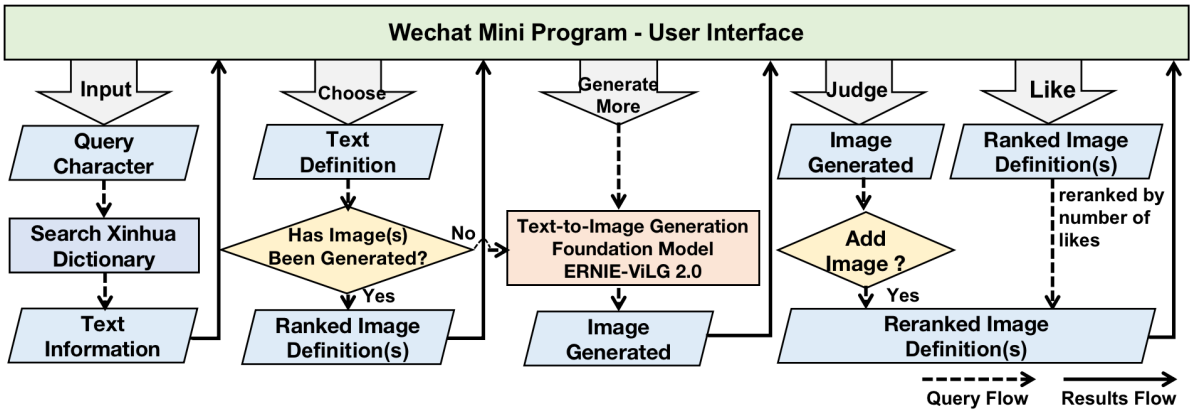


Figure 1: System Framework and Workflow

(a) Generated images of Definition 2 of character "Yuan" queried by description only.

(b) Generated images of Definition 2 of character "Yuan" queried by both description and sample words.

Figure 2: Image Generation with Different Painting Objects of Definition 2 of Character "Yuan"

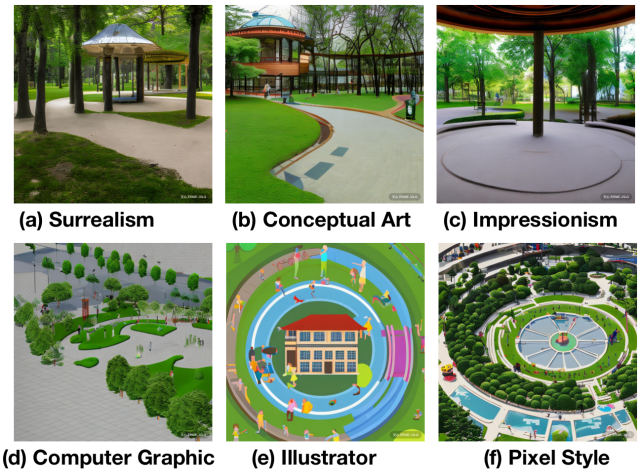


Figure 3: Image Generation with Different Painting Styles of Definition 2 of Character "Yuan"

2.3 Collaborative Learning

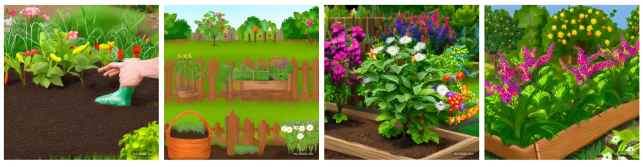
Based on the text-to-image generation results of the foundation model, the system supports learners to collaboratively refine the image definitions. As shown in Figure 2(b) and Figure 4, the image definitions of three text definitions of character "Yuan" are equally important to learners, which demands a high cognitive load to understand them all. To improve the learners' understandings of character definitions, the system encourages learners to vote for the most suitable images based on their understandings, and add new images as candidates when none of the generated images are favored. Ideally, the image with the most votes would be displayed at the top of the list on the user interface and would be considered as the most appropriate image definition to the text definition based on collective knowledge.

The voting and adding image processes require learners to review the text definition carefully and figure out the key semantic meaning of the AI-generated image. Comparing the similarity between the text and image definitions in mind, learners could strengthen the comprehension of the character via verbal and visual dual-channels before making the rational voting decision. When new learners searching the same character, the previous work would support them understanding the text definitions accompanied with most relevant images ranked by others' votes. Meanwhile, new learners could also be inspired to progressively make contributions to the system and work collectively to develop the community knowledge.

3. USER INTERFACE

The user interface of the system is based on WeChat mini program which is a mobile application accessed through WeChat, the most popular social software in China, without extra downloading. Learners could operate it on mobile devices wherever in formal or informal learning environment. As shown in Figure 5, learners could input the query Chinese character in the search box and click on the search button. The system would then return basic information with multiple text definitions of the query character.

After that, as shown in Figure 6, learners could click on each text definition to show the corresponding image definitions, where the generated images are ranked by the number of likes voted by other learners. It requires learners to browse the generated images from the top-ranked to the bottom,



(a) Generated images of Definition 1 of character "Yuan" queried by both description and sample words in realistic style

(b) Generated images of Definition 3 of character "Yuan" queried by both description and sample words in realistic style

Figure 4: Generated Images of Definition 1 and 3 of Character "Yuan"

(a) Input Query Character (b) Search Results

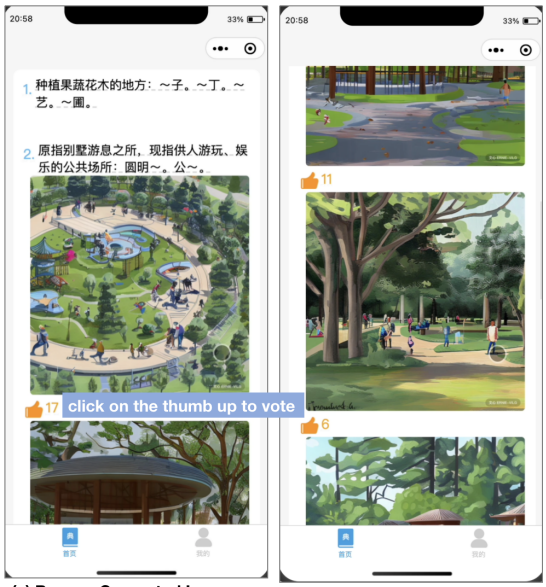
Figure 5: User Interface of Character Querying

and make their own decisions to vote for the appropriate images by clicking on the thumb up button.

When none of the generated images suitable for the text definition, learners could choose to generate new image by clicking on the "generate my image" button at the bottom of the list, as shown in Figure 7. It takes around 10-20 seconds to generate an image with resolution of 1024×1024 pixels. Before adding to the list, a popup would ask for learner's confirmation, which expects the learner to review the text definition and make a deliberate decision for the image definition.

4. CONCLUSION AND FUTURE WORK

We propose a multimodal language learning system for Chinese character with the help of text-to-image generation foundation model ERNIE-ViLG2.0. Based on the text-to-image generation results, learners could help to improve others' understandings of Chinese character definitions by voting and adding images to re-rank the images' display order. Consequently, learners could benefit from the top-ranked images for each character's text definition and improve the cognition through both verbal and visual channel.



(a) Browse Generated Images (b) Vote for Appropriate Images

Figure 6: User Interface of Images Browsing and Voting

(a) Generate New Image (b) Confirm to Add the Image

Figure 7: User Interface of Images Generation and Adding

In the future work, to estimate the effectiveness of the system, we plan to design and conduct experiments by inviting entry-level Chinese learners to evaluate their learning achievement and attitudes towards the generated images and the voting system. Especially, it is also worth to investigate the effectiveness of various style images and how they provide improvement in the learning process. Besides, consid-

ering about the quality of generated images, the trust of the voting system requires further supervisions to correct typical mistakes from beginning learners and avoid unfriendly attacks.

Additionally, more flexible functions could be added to the built system. For example, in addition to "like" button, "dislike" could also be an option to express learner's opinions on the image. Further, to deepen learners' understandings, it also welcomes learners to make text comments on the image and leave nicknames and avatars to improve community awareness. Besides, since foundation models are pre-trained on large-scale data by black-box method, it is also necessary to require interventions to avoid risks of algorithm biases and intellectual property issues.

Furthermore, the multimodal language learning system could also be transferred to other languages learning with the similar mechanisms of text-to-image generation and learners' collaboration. Additionally, foundation models for AI generation are also powerful on text-to-text generation, image-to-text generation, image modification, etc. It would be interesting to investigate more possibilities of interaction and integration with AI-generated content and learner-generated content.

5. ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500; in part by National Natural Science Foundation of China (Grant No. 62077006).

6. REFERENCES

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv preprint arXiv:2210.15257*, 2022.

[4] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[5] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[6] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[7] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[8] R. Oxford and D. Crookall. Vocabulary learning: A critical analysis of techniques. *TESL Canada journal*, pages 09–30, 1990.

[9] A. Paivio. *Mental representations: A dual coding approach*. Oxford University Press, 1990.

[10] J. L. Plass, D. M. Chun, R. E. Mayer, and D. Leutner. Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of educational psychology*, 90(1):25, 1998.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.