# An Intelligent Multimodal Dictionary for Chinese Character Learning

Jinglei Yu, Jiachen Song, Penghe Chen, and Yu Lu(✉)

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

**Abstract.** Chinese character learning is difficult, as the character's definitions in dictionary are simple but abstract. The image representations of Chinese character's definitions are easy to understand and helpful to remember. To assist learning Chinese character and understanding definitions, we design an intelligent dictionary which supports text and image of printed character as input and text, image and video as output modes. Particularly, users could query each definition in text to obtain the corresponding image definition via the designed cross-modal retrieval mechanism. Besides, we also build the image database of character evolving process as well as the video databases of micro-lectures for extended learning. A mobile version of the dictionary has been developed, which supports the multimodal query and output information for the individual Chinese character.

**Keywords:** Chinese character learning · Cross-modal retrieval · Multimodal dictionary

## 1 Introduction

Chinese character learning is a challenging task for learners, as it is hard to recognize single character, understand its multiple meanings, make appropriate phrases and form long-term memories. Chinese dictionary is an efficient and useful tool to learn Chinese characters, mainly containing pinyin, glyph information and multiple definitions. Pinyin refers to the character's pronunciation and glyph typically includes character's structure, radical (semantic or phonetic component), and number and sequence of strokes. Definition is the statement of character's meanings in different context using simple but abstract description.

By leveraging the current Chinese dictionary, learners could obtain the necessary information of the individual character as well as the commonly used phrases. However, it is still difficult for learners to get a quick understanding and form long-term retention of all the character's information, especially the character's multiple and ambiguous definitions. The previous studies show that images are appropriate for representing abstract scenario and unusual objects [5]. According to the dual coding theory [4], the verbal and visual dual representation

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13356, pp. 79–83, 2022.
https://doi.org/10.1007/978-3-031-11647-6_13

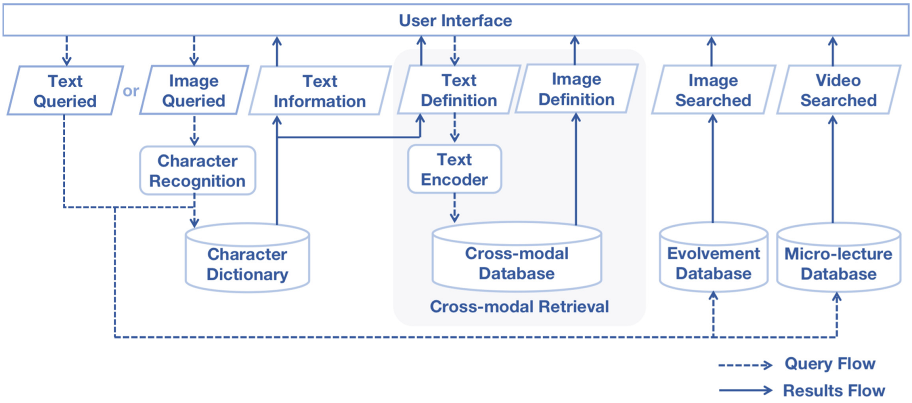


**Fig. 1.** The simplified block diagram of the designed dictionary.

could enhance learner's memory, especially when word and image are strongly associated with each other [7]. Hence, we design and implement an intelligent Chinese dictionary featured with multimodal input and output. Each definition of the character could be queried to show its image representation. Besides, since most Chinese characters have their unique historical evolving processes, where the original script is pictorial and may reflect the visual meaning from the perspective of character formation [8], we also provide characters' evolving process images and correspondent micro-lectures videos for extended learning.

## 2 Dictionary Design

Figure 1 illustrates the block diagram of the dictionary. The dictionary supports two types of retrieval functions, namely multimodal information search and cross-modal retrieval.

### 2.1 Multimodal Information Search

In Fig. 1, the whole framework except the cross-modal retrieval part is the multimodal information search part. The input supports either typing a character or uploading an image of the printed character, where the optical character recognition (OCR) service is used to extract the queried character from the image. Based on the queried character, the system requests the online API of Xinhua dictionary, which is the most popular Chinese dictionary, for the character's basic text information, including pinyin, glyph information and definitions. For the character evolving process, we build a dedicated database to store the image collections of characters in five chronologically formed scripts from calligraphy works. We also build another database for 15 exemplary characters by manually recording 1–2 min micro-lectures for each character to further explain their glyphs and definitions from the historical perspective.

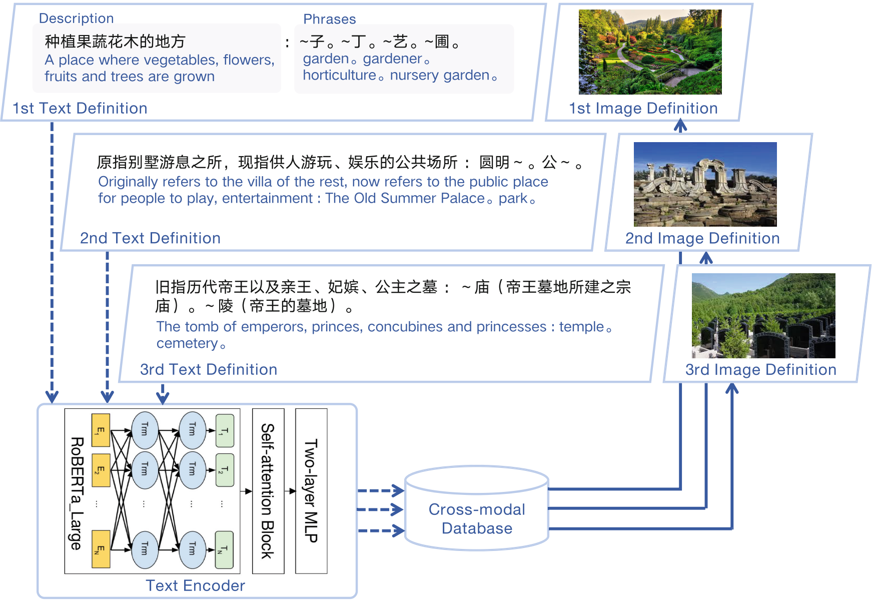

Description
种植果蔬花木的地方
A place where vegetables, flowers, fruits and trees are grown
Phrases
: ~子。~丁。~艺。~圃。
garden。gardener。
horticulture。nursery garden。
1st Text Definition
1st Image Definition
原指别墅游息之所，现指供人游玩、娱乐的公共场所 ： 圆明 ~ 。公 ~ 。
Originally refers to the villa of the rest, now refers to the public place for people to play, entertainment : The Old Summer Palace。park。
2nd Text Definition
2nd Image Definition
旧指历代帝王以及亲王、妃嫔、公主之墓 ： ~ 庙（帝王墓地所建之宗庙）。~ 陵（帝王的墓地）。
The tomb of emperors, princes, concubines and princesses : temple。cemetery。
3rd Text Definition
3rd Image Definition
RoBERTa_Large
E1
E2
EN
Tm
T1
T2
TN
Self-attention Block
Two-layer MLP
Text Encoder
Cross-modal Database

**Fig. 2.** The simplified block diagram and its query results of cross-model retrieval.

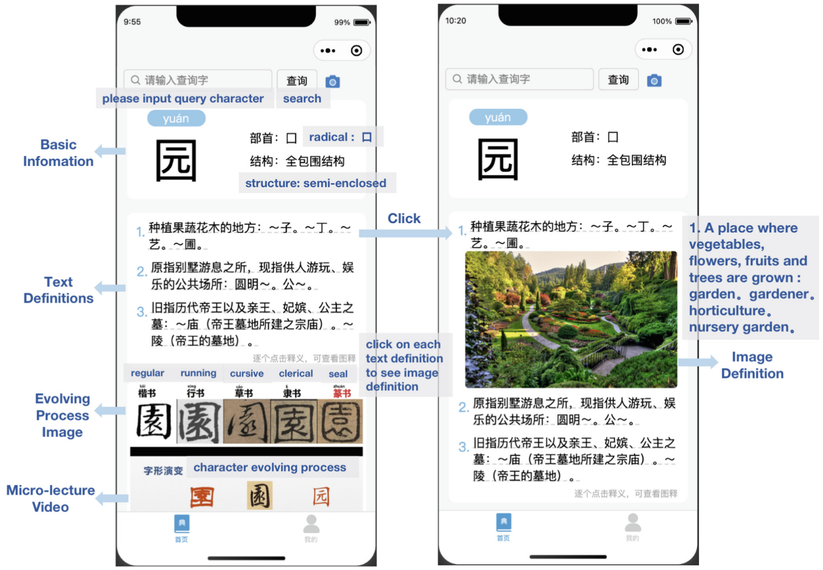
## 2.2 Cross-modal Retrieval

Based on the multimodal information search results, the system supports learner clicking on each text definition to query its image definition. As shown in Fig. 2, the queried Chinese character is "Yuan" and each definition could be split into a description and example phrases. Each of them is a query string. The cross-model retrieval mechanism firstly extracts each text's feature by text encoder and searches the image with maximum cosine similarity in the cross-modal database. After that, the image from text-image pairs with the maximum similarity would be selected as the image definition. The image features are extracted and stored in the database in advance.

In practice, the text and image features are extracted by encoders from large-scale multimodal pre-trained model BriVL [1,2]. Specifically, BriVL is a two-tower training framework consisting of two replaceable text and image encoders, which are connected by InfoNCE loss in the training process. After the pre-training, the two encoders could work independently and provide APIs that we utilized. The image encoder is based on Efficient-Net_B7 [6] and the text encoder is based on the Chinese pre-trained version of RoBERTa_Large [3]. Both of them are followed by self-attention block and multi-layer perception (MLP) block to project features to the same cross-modal space. The pre-trained model is learned from 650 million image-text pairs crawled from web. The wide coverage of topics and scenarios is rationally enough for our retrieval design.

### 2.3 User Interface

A mobile version of the dictionary has been developed as the user interface, which could be accessed without downloading, as shown in Fig. 3. Users could look up the dictionary in either formal or informal learning environment. The input could be either typing or simply uploading a picture of the printed character for both native and non-native speakers. To reduce the cognitive load and deepen the impression of the definition, learner could choose to click on each text definition and show its image definition. For extended learning, the character evolving process and micro-lectures are provided. From the perspective of character formation, micro-lectures analyze character's glyph, original meanings, shape changes during the historical process and its current usages in phrases.



1. 种植果蔬花木的地方：～子。～丁。～艺。～圃。
2. 原指别墅游息之所，现指供人游玩、娱乐的公共场所：圆明～。公～。
3. 旧指历代帝王以及亲王、妃嫔、公主之墓：～庙（帝王墓地所建之宗庙）。～陵（帝王的墓地）。

逐个点击释义，可查看图释

1. A place where vegetables, flowers, fruits and trees are grown : garden。gardener。horticulture。nursery garden。

**Fig. 3.** The user interface of the dictionary.

## 3 Conclusion and Future Work

By leveraging the multimodal pre-trained model, we design and implement the intelligent Chinese dictionary with interactive image representation for each definition to smooth the learners' path to acquire Chinese characters. Besides, 15 exemplary characters' evolving processes and micro-lectures are provided and implemented on the mobile version, which would be constantly enlarged to cover more basic Chinese characters. For the future work, the text-to-image retrieval recall of the encoders needs further improvement by updating state-of-art single-modal encoders for BriVL and fine-tuning the cross-modal framework. Besides, user's feedback mechanism would also be useful to correct inaccurate image definitions and collect bad cases for model fine-tuning. Additionally, considering cloud storage load, images should be collected from online resources and only image URL-feature pairs are stored locally with the regularly validation checking. We are currently deploying the dictionary to serve the school students and teachers, and the usability study is also in plan.

**Acknowledgements.** This research is supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (No. 62077006, 62177009).

## References

1. Fei, N., et al.: WenLan 2.0: make AI imagine via a multimodal foundation model. arXiv preprint arXiv:2110.14378 (2021)
2. Huo, Y., et al.: WenLan: bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561 (2021)
3. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
4. Paivio, A.: Mental Representations: A Dual Coding Approach. Oxford University Press, Oxford (1990)
5. Szczepaniak, R., Lew, R.: The role of imagery in dictionaries of idioms. Appl. Linguis. **32**(3), 323–347 (2011)
6. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
7. Underwood, J.: HyperCard and interactive video. Calico J. 7–20 (1989)
8. Yu, J., Song, J., Lu, Yu., Yu, S.: Back to the origin: an intelligent system for learning chinese characters. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12749, pp. 457–461. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78270-2_81