MChIRC: A Multimodal Benchmark for Chinese Idiom Reading Comprehension

Tongguan Wang[*1, 2, 3, 4], **Mingmin Wu**[*4], **Guixin Su**[4], **Dongyu Su**[4], **Yuxue Hu**[1, 2, 3, 4], **Zhongqiang Huang**[4], **Ying Sha**[†1, 2, 3, 4]

[1]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
[3]Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China
[4]College of Informatics, Huazhong Agricultural University, Wuhan, China
{wang_tg, wmm_nlp, cometsue, su_dy, hzq}@webmail.hzau.edu.cn, {hyx,shaying}@mail.hzau.edu.cn

Abstract

The performance of various tasks of natural language processing has greatly improved with the emergence of large language models. However, there is still much room for improvement in understanding certain specific linguistic phenomena, such as Chinese idioms, which are usually composed of four characters. Chinese idioms are difficult to understand due to semantic gaps between their literal and actual meanings. Researchers have proposed the Chinese idiom reading comprehension task to examine the ability of large language models to represent and understand Chinese idioms. The task requires choosing the correct Chinese idiom from a list of candidates to complete the sentence. The current research mainly focuses on text-based idiom comprehension. Nevertheless, there are many idiom application scenarios that combine images and text, and we believe that the corresponding images are beneficial for the model's understanding of the idioms. Therefore, to address the above problems, we first construct a large-scale **M**ultimodal **Ch**inese **I**diom **R**eading **C**omprehension dataset (**MChIRC**), which contains a total of 44,433 image-text pairs covering 2,926 idioms. Then, we propose a **D**ual-**C**ontrastive **I**diom **G**raph **N**etwork (**DCIGN**), which employs a dual-contrastive learning module to align the text and image features corresponding to the same Chinese idiom at both coarse and fine levels, while utilizing a graph structure to capture the semantic relationships between idiom candidates. Finally, we use a cross-attention module to fuse multimodal features with graph features of candidate idioms to predict correct answers. The authoritativeness of MChIRC and the effectiveness of DCIGN are demonstrated through a variety of experiments, which provides a new benchmark for the multimodal Chinese idiom reading comprehension task.

Introduction

Chinese idioms are a special form of linguistic expression, usually consisting of four Chinese characters, such as "绝处逢生 (Rescued from desperation)", as shown in Figure 1. The challenge in understanding Chinese idioms is the inconsistency between the literal meaning and the metaphorical meaning, which usually requires some background knowledge, such as history. Therefore, obtaining accurate representation and understanding of idioms is crucial for downstream tasks such as machine translation (Li et al. 2024), image generation (Ali et al. 2024), and image-text matching (Yosef, Bitton, and Shahaf 2023).

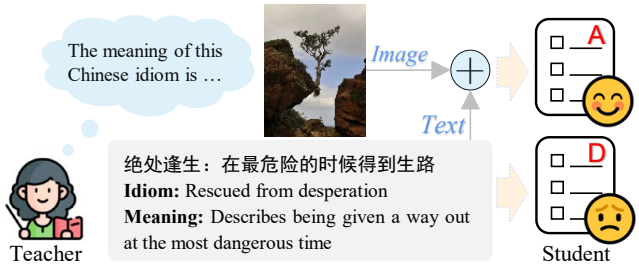


Figure 1: An intriguing phenomenon is observed during the classroom instruction of idioms.

Therefore, Zheng, Huang, and Sun (2019) proposed a cloze task for idioms and established the widely used Chinese idiom completion dataset, ChID. The cloze task involving idioms typically requires the model to select the most suitable idioms from the set of candidate idioms to be placed in the blanks of the sentence, which requires the model to maximize the understanding of the special meanings of different idioms in different contexts.

The current methods are mainly based on deep learning models and have achieved certain performances. Tan and Jiang (2021) proposed a BERT-based dual embedding model and a two-stage model with context pooling and fine-tuning to improve the accuracy of idiom prediction. Wang et al. (2020) introduced an attribute focus mechanism to correct idiomatic misuse. Sha et al. (2023) proposed a prompt-based representation individual enhancement approach to learn idiomatic metaphorical meaning. Wu et al. (2024a) addressed the problem of inconsistency between idiomatic metaphor and context through multi-semantic connectivity and contrastive learning, achieving state-of-the-art performance with their method in the ChID dataset. However, all these existing methods consider only textual features. The rapid development of social media has led to a diverse range of data sources for idioms. How to effectively utilize heterogeneous information such as images corresponding to idioms to assist the model in accurately understanding the id-

[*]These authors contributed equally.
[†]Corresponding author.

Dataset	Sentence	Number of idioms	Average words	Modality	
				Text	Image
People Daily (Cui et al. 2016)	876,710	248,160	39	✓	✗
Children's Fairy Tale (Cui et al. 2016)	3,599	-	20	✓	✗
CMRC-2017 (Cui et al. 2018)	364,295	94,352	17.4	✓	✗
CIBB (Shao et al. 2018)	1,194	50	9.81	✓	✗
CCT (Jiang et al. 2018)	108,987	7,395	27.37	✓	✗
ChID (Zheng, Huang, and Sun 2019)	580,807	3,848	100	✓	✗
ChIdSyn (Tan and Jiang 2021)	21,000	8,125	97.66	✓	✗
CIP (Qiang et al. 2023)	115,529	8,421	36.55	✓	✗
CIDT (Wu et al. 2024a)	3,600	3,848	22.12	✓	✗
MChIRC(Ours)	44,433	2,926	15.42	✓	✓

Table 1: Summary of datasets of Chinese idioms datasets.

ioms is an urgent issue to be considered.

In reality, images and words related to idioms often appear simultaneously, which facilitates people's understanding of idioms. In a real teaching case as shown in Figure 1, if only words are used to describe the meaning of the idiom "绝处逢生" (Rescued from desperation), students may lack intuitive feelings to understand "what is the most dangerous situation?" and "how to get out?". If a picture is added, and the picture shows "a tree growing on the edge of a cliff", it will be easier for students to understand the meaning and application scenarios of this idiom. Accordingly, these students are more likely to get an "A" in the test (Wong et al. 2010; Pintado and Fajardo 2021; Zhang 2021). Inspired by this phenomenon, the Chinese idiom reading comprehension task must also consider multimodal scenarios. In addition, the context of Chinese idioms in existing datasets is generally longer, while the context of Chinese idioms in actual scenarios is often shorter.

To address the above issues, we build, to the best of our knowledge, the first multimodal Chinese idiom reading comprehension dataset (MChIRC[1]). We crawl numerous images from Baidu[2] and Sogou[3]. After manual annotation, we collect 44,433 image-text pairs covering 2,926 idioms. The average length of sentences in the dataset is 15.42, which is more consistent with the context length of Chinese idioms in actual scenarios (Wang et al. 2021). The content of the MChIRC dataset covers multiple fields such as news, comics, animation, and product information.

Based on the constructed multimodal dataset, we propose a dual-contrastive idiom graph network (DCIGN), which employs a dual-contrastive learning module to align the text and image features corresponding to the same Chinese idiom at both coarse and fine levels, thus improving the model's ability to recognize positive and negative idiom text-image pairs. In addition, the graph structure is utilized to learn the semantic relationship between the candidate idioms and further distinguish the nuances between different candidate idioms. Finally, we use a cross-attention module to fuse multimodal features with graph features of candidate idioms to predict correct answers. Extensive experiments conducted on the MChIRC dataset demonstrate the effectiveness of our proposed method, achieving an average accuracy of 73% in the four test sets.

The main contributions are as follows:

- We construct the first multimodal Chinese idiom dataset, MChIRC, which is more aligned with real-world scenarios. The average length of the sentences is only 15.42.
- We propose a dual-contrastive idiom graph network, DCIGN, which employs a dual-contrastive learning module to align the text and image features corresponding to the same Chinese idiom at both coarse and fine levels, while a graph structure is utilized to learn the semantic relationships among idiom candidates.
- Experimental results demonstrate that DCIGN achieves promising performance on the MChIRC dataset, which proves the rational utilization of visual information and provides a novel benchmark for Chinese idiom research.

Related Work

Datasets

As a unique linguistic phenomenon in Chinese, Chinese idioms have attracted numerous scholars to conduct extensive research. We summarize and analyze the Chinese idiom datasets in recent years, as presented in Table 1. The Chinese idiom dataset is currently primarily applied to four types of tasks, which are used for the Chinese idiom reading comprehension task (Cui et al. 2016, 2018; Jiang et al. 2018; Zheng, Huang, and Sun 2019; Wu et al. 2024a), the Chinese idiom translation task (Shao et al. 2018), the Chinese idiom embedding task (Qiang et al. 2023), and the Chinese idiom rewriting task (Tan and Jiang 2021). In summary, these Chinese idiom datasets have provided abundant resources for the research of Chinese idioms.

However, we find that the average words of sentences in these datasets are usually large, which usually means that they contain ample contextual information. In the real-world application of Chinese idioms, news headlines often use concise sentences that include idioms to attract attention, which means that context information for Chinese idioms is sparse, making the prediction of correct idioms more challenging (Knietaite et al. 2024). It is particularly noteworthy

[1]https://github.com/Aichiniuroumian/MChIRC.

[2]https://image.baidu.com/

[3]https://pic.sogou.com/

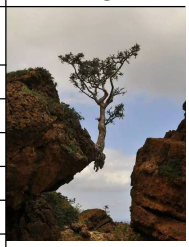
Text & Candidate	Image
#idiom# 的树, 生命真的是强大啊! #idiom# trees, life is really powerful!	
NO.1 否极泰来 Adversity leads to prosperity	
NO.2 引而不发 Draw the bow without shooting	
NO.3 峰回路转 The twists and turns	
NO.4 颠扑不破 Indisputable	
NO.5 绝处逢生 Rescued from desperation ✓	
NO.6 指手划脚 Point and gesture with one's fingers	*正确选项是 NO.5* *The correct choice is NO.5*
NO.7 起死回生 Bring the dying back to life	

Figure 2: An example of the MChIRC dataset.

that these datasets only utilize textual information, neglecting the potential of image information to assist in task learning, as demonstrated by other multimodal datasets (Jin et al. 2017; Cai, Cai, and Wan 2019; Qi et al. 2023).

Chinese Idiom Reading Comprehension

The Chinese idiom comprehension task aims to make models understand the exact meaning of idioms. Cui et al. (2016, 2018) used the complete representation of the queries and the "Overly Attentive Reader" to solve reading comprehension tasks. Jiang et al. (2018) used background knowledge to improve the accuracy of the cloze task. Zheng, Huang, and Sun (2019) proposed three baselines on the ChID dataset: LM, AR, and SAR. Subsequent studies, such as those using attention mechanisms by Long et al. (2020) and Wang et al. (2020), a two-stage model by Tan and Jiang (2021), and interpretative approaches by Dai et al. (2023) and Sha et al. (2023), have significantly improved machine understanding of Chinese idioms. In addition, Wu et al. (2024a) addressed the problem of inconsistency between idiom literal and metaphorical meanings through multi-semantic contrastive learning, achieving state-of-the-art performance on the ChID dataset with an accuracy of 96.8% on the test set.

Nevertheless, to the best of our knowledge, there is a gap in the study of multimodal Chinese idioms. To address this, we construct MChIRC, a more relevant and challenging multimodal Chinese idioms dataset tailored for real-world scenarios, with an average sentence length of only 15.42 words. We also propose a dual-contrastive idiom graph network, DCIGN, which integrates both text and image features for the first time to tackle the Chinese idiom reading comprehension task. Our method provides a novel perspective for research on Chinese idioms.

Task Definition

The objective of the multimodal Chinese idiom cloze task is to select the most appropriate idiom from a set of seven candidates, based on a given text segment and an accompanying image. The text is denoted as $T = \{w_1, w_2, ..., [MASK], ..., w_i, w_n\}$, where each character w_i represents a Chinese character, and the position of the idiom is marked with $[MASK]$. The image is denoted as $I = \{image\}$. The candidate idiom set is denoted as $C = \{c_1, ..., c_i, ..., c_k\}$, comprises six distractor idioms and one correct idiom. We present a data sample in Figure 2. The text reads "#idiom#的树, 生命真的是强大啊! (#idiom# trees, life is really powerful!)". The image visually depicts a tree growing on cliffs. Therefore, in this example, the correct option for "#idiom#" can be identified as "NO.5 绝处逢生 (NO.5 Rescued from desperation)", which perfectly captures the essence of the text and image.

DCIGN Model

Model Overview

We propose a dual-contrastive idiom graph network (DCIGN) for multimodal Chinese idiom reading. As shown in Figure 3, DCIGN consists of five main components: Unimodal Feature Extraction, Fine-Grained Feature Extraction, Dual-Contrastive Learning from Coarse to Fine, Graph Relationship Modeling for Idiom Candidates, and a Predict Module.

Unimodal Feature Extraction

The input text for the network is denoted as $T = \{w_1, w_2, ..., [MASK], ..., w_i, w_n\}$, while the input image for the network is a corresponding image of the idiom, denoted as $I = \{image\}$. To obtain the most primitive features from different modalities, similar to most multimodal approaches (Huang et al. 2023), we utilize a pre-trained BERT[4] to extract textual features and a pre-trained DeiT[5] to extract image features. The formulations are detailed as follows:

$$T_o = BERT([CLS], w_1, w_2, \ldots, [MASK], \ldots, w_i, w_n, [SEP]), \quad (1)$$

$$V_o = DeiT(I), \quad (2)$$

where T_o represents the extracted primitive text features. V_o represents the extracted primitive image features. The special marks $[CLS]$ and $[SEP]$ are boundary marks used to guide and terminate the input. The $[MASK]$ refers to idioms that are masked out.

Fine-grained Feature Extraction

Due to the presence of redundant or interfering information in the originally extracted features, the model's ability to understand text or image information is affected to some extent. Therefore, we use self-attention to refine the key information in the extracted primitive features and ignore the information that is not important for understanding idioms, respectively. The formulations are detailed as follows:

$$\tilde{T}_o = LN\left(SA\left(Q_{T_o}, K_{T_o}, V_{T_o}\right) + T_o\right), \quad (3)$$

$$\tilde{V}_o = LN\left(SA\left(Q_{V_o}, K_{V_o}, V_{V_o}\right) + V_o\right), \quad (4)$$

where $\tilde{T}_o$ represents refined text features. $\tilde{V}_o$ represents refined image features. LN stands for Layer Normalization, SA stands for Self-Attention.

Then, we fuse the extracted refined features with the extracted original features to obtain the final multimodal features F_{fuse} as follows:

$$F_{fuse} = T_o + \tilde{T}_o + V_o + \tilde{V}_o. \quad (5)$$

[4]https://huggingface.co/hfl/chinese-bert-wwm

[5]https://huggingface.co/facebook/deit-base-patch16-224

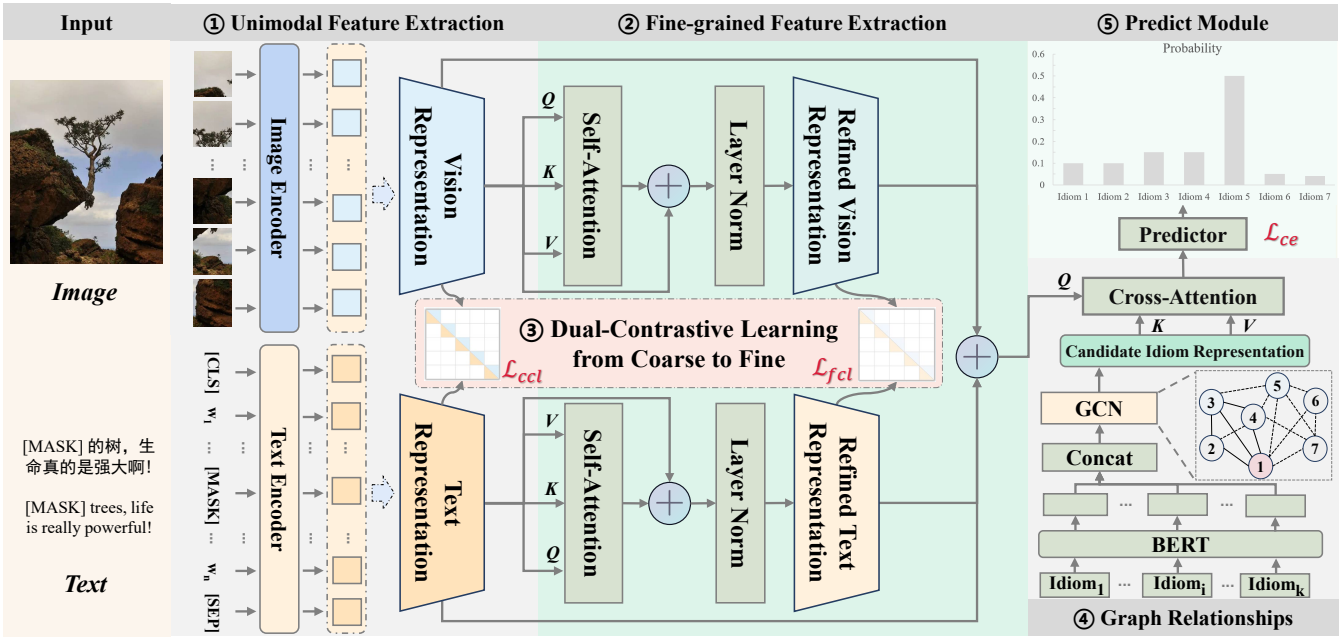


Figure 3: Dual-Contrastive Idiom Graph Network for Multimodal Chinese Idiom Reading.

Dual-Contrastive Learning from Coarse to Fine

Contrastive learning has been achieved with impressive results across tasks (Zhang et al. 2024; Su et al. 2024; Hu et al. 2024). The primitive and refined features focus on different regions in the text and images, potentially causing fine-grained errors despite coarse-grained alignment. Therefore, to improve the alignment of the model at different levels of modal features, we use a dual-contrastive learning module to align multimodal features.

Firstly, we use coarse contrastive learning (CCL) to align the primitive text and image features. Coarse contrastive learning helps the model align features that are roughly different. Then, to further refine and optimize the feature representation and enhance the model's ability to distinguish subtle semantic differences, we use fine contrastive learning (FCL) to process the finely extracted text and image features. Based on coarse contrastive learning, fine contrastive learning further adjusts and optimizes the layout of the feature space, so that the text-image features of the same idiom are more precisely aligned, while the text-image features of different idioms are more clearly distinguished.

$$\mathcal{L}_{ccl} = -\sum_{i=1}^{N} log \frac{e^{sim\left(T_0^i, V_0^i\right)/\tau}}{\sum_{j=1}^{N} e^{sim\left(T_0^i, V_0^j\right)/\tau}}, \tag{6}$$

$$\mathcal{L}_{fcl} = -\sum_{i=1}^{N} log \frac{e^{sim\left(\tilde{T}_0^i, \tilde{V}_0^i\right)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(\tilde{T}_0^i, \tilde{V}_0^j\right)/\tau}}, \tag{7}$$

where N represents the batch size, (T_0^i, V_0^i) stands for the i-th sample comes from two different modal features. sim is the cosine similarity. τ is the temperature hyper-parameter.

Coarse contrastive learning and fine contrastive learning collaborate with each other, which not only aligns the text image features of the same idiom at two levels but also enhances the model's sensitivity to small variations, enabling it to make more accurate and fine-grained judgments when faced with complex and subtle multimodal idiom reading comprehension cloze tasks.

Graph Relationship Modeling for Idiom

We also use pre-trained BERT to extract the features of the idioms in the candidate set. The formulations are as follows:

$$C_k = BERT(Idiom_k), \tag{8}$$

where k stands for the k-th idiom, $k \in [1, 7]$.

After obtaining the feature representation of each candidate idiom C_k, we concatenate these features to get the feature representation C of the candidate idiom.

$$C = Concat(C_1, C_2, \ldots, C_k). \tag{9}$$

The idioms among the candidates exhibit semantic associations. Taking the idiom candidate set in the test set as an example, it comprises three idioms most similar to the correct option and three idioms randomly selected. We consider each candidate idiom as a node and the relationship between them as edges. The representation of candidate idioms is not solely determined by their individual characteristics, it can also aggregate information from adjacent idiom nodes through a graph structure, thus obtaining a richer and more distinctive representation. Therefore, we introduce a graph structure to capture the semantic representations between idioms. The feature update process for each node c at layer l is as follows:

$$h_c^{(l+1)} = ReLU\left(\sum_{u \in \mathcal{N}(c)} \frac{1}{p_{uc}} W^{(l)} h_u^{(l)}\right), \tag{10}$$

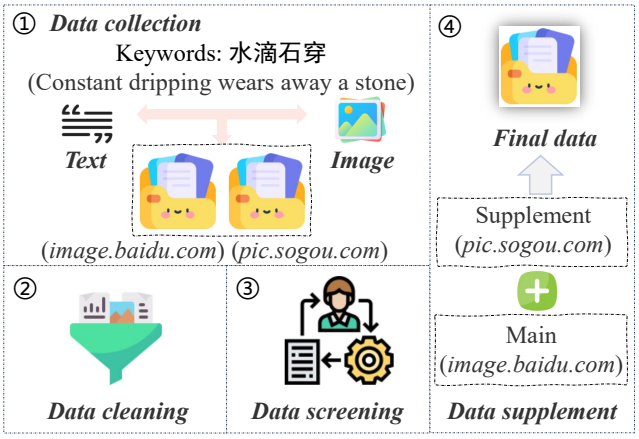


Figure 4: Flowchart of MChIRC dataset construction.

where $h_u^{(l)}$ represents the feature vector of the neighbors u of idiom node c at layer l. $\mathcal{N}(c)$ denotes the set of neighboring idiom nodes of idiom node c. p_{uc} is the normalization coefficient, which is used to control the influence of the features of neighboring idiom nodes on the features of the current idiom node. $W^{(l)}$ is the weight matrix at layer l.

After the graph convolution layer, we concatenate the representations of each node to obtain the overall representation of the candidate idiom, denoted by h_v.

Prediction Module

In order to capture the intention and context of the query more accurately. We use fused multimodal features as query Q and candidate idioms that have been through the graph structure as key K, value V. The two features are fused using the cross-attention module, and the final representation of the fusion feature $\tilde{M}$ is obtained.

$$\tilde{M} = CA(Q_{F_{fuse}}, K_{h_v}, V_{h_v}). \tag{11}$$

We feed the fusion feature representation $\tilde{M}$ into a linear layer, which then undergoes a softmax function for probabilistic prediction of the seven idioms in the candidate set.

$$P(Idiom_k|T, I) = Softmax(W\tilde{M} + b), \tag{12}$$

where $Idiom_k$ stands for the k-th idiom in the candidate set.

We minimize the cross-entropy loss function to calculate the difference between the predicted probability distribution and the true distribution for the correct idiom, as follows:

$$\mathcal{L}_{ce} = -\sum_{k=1}^{7} C_g log(P(Idiom_k|T, I)), \tag{13}$$

where C_g stands for the one-hot label distribution for the correct idiom.

Using joint dual-contrastive learning loss, we end up with a total loss as shown as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha\mathcal{L}_{ccl} + \beta\mathcal{L}_{fcl}, \tag{14}$$

where $\alpha + \beta = 1$, α and β are the weights of the coarse contrastive learning loss and the fine contrastive learning loss, respectively.

	Train	Dev	Test/Sim	Out	Total
Idiom	2,926	2,456	2,881	2,926	2,926
Pair	29,054	4,151	8,302	2,926	44,433

Table 2: Division of the MChIRC dataset. The data in the Test set and the Sim set are identical except for the different composition of the set of candidate idioms.

Experiments

Datasets

We present an overview of the construction process for the MChIRC dataset, as illustrated in Figure 4. In general terms, the process is primarily divided into four stages: data collection, data cleaning, data screening, and data supplement. All data are used for scientific research only. We promise that the data collected will not be used for non-commercial or profit-making purposes and strictly respect this dataset's copyright.

We evaluate all methods using the MChIRC multimodal idiom dataset. Initially, we extract one sample from each idiom as the Out set, which is the most challenging. The remaining samples are divided into a training, validation, and test set in a 7: 1: 2 ratio using stratified sampling to ensure as much balance as possible between interclass and intraclass data. To further test the robustness of the models, we construct the Sim set following Zheng, Huang, and Sun (2019). The set of candidate idioms for the Sim set consists of the six idioms that are semantically most similar to the correct idiom. The final dividend dataset is shown in Table 2.

Baselines

We compare six text modality methods, three image modality methods, and four classic multimodal methods as baselines for comparison. These methods are described in detail below.

- **BERT-WWM** (Cui et al. 2021): An upgraded version of the BERT model that uses whole word masking.
- **RoBERTa** (Cui et al. 2021): An improved BERT employs dynamic masking, more pre-training data, and extended training time.
- **MacBERT** (Cui et al. 2021): An improved version of BERT, specifically designed for Chinese, introduces the pre-training task of MLM as correction (Mac).
- **PRIEM** (Sha et al. 2023): Fusing the definitions of idioms through the prompt method and then using orthogonal projection to distinguish idioms' representations.
- **RISCF** (Wu et al. 2024b): The accuracy of idiom completion tests has been enhanced through semantic contrast learning and an anti-interference cross-attention module.
- **MSCLM** (Wu et al. 2024a): The model addresses the issues of metaphorical inconsistency and contextual inconsistency in idioms. By using metaphor contrastive learning and multi-semantic cross-attention modules.

Modality	Method	Dev-Acc	Test-Acc	Sim-Acc	Out-Acc	Avg-Acc
Text	BERT-WWM (Cui et al. 2021)	63.43	60.40	59.50	49.79	58.28
	RoBERTa (Cui et al. 2021)	64.49	61.85	60.50	50.72	59.39
	macBERT (Cui et al. 2021)	64.27	61.43	60.20	50.31	59.05
	PRIEM (Sha et al. 2023)	65.94	64.24	71.30	47.13	62.15
	RISCF (Wu et al. 2024b)	67.02	66.12	66.65	46.82	61.65
	MSCLM (Wu et al. 2024a)	71.19	69.56	67.89	51.95	65.15
Image	VGG-16	45.70	44.05	45.27	33.36	42.10
	CLIP (Radford et al. 2021)	40.66	38.63	42.57	26.35	37.05
	DeiT (Touvron et al. 2021)	55.46	54.01	54.81	40.64	51.23
Multimodal	MSCA† (Huang et al. 2023)	68.63	69.15	68.01	55.04	65.21
	Multi-view CLIP† (Qin et al. 2023)	59.05	61.05	51.04	49.24	55.10
	DivE† (Kim, Kim, and Kwak 2023)	74.24	71.56	66.71	56.03	67.13
	CLTL† (Wang and Markov 2024)	65.55	64.05	60.24	50.75	60.15
	DCIGN (Ours)	**77.26**	**77.16**	**77.34**	**60.25**	**73.00**

Table 3: Comparison results (%) with baseline models on the MChIRC dataset. "†" stands for rerunning these methods.

- **VGG-16** (Simonyan and Zisserman 2015): Enhancing model representation by stacking multiple smaller convolutional and pooling layers.
- **CLIP** (Radford et al. 2021): A multimodal pretraining model designed to combine text and image information for robust cross-modal understanding and generation.
- **DeiT** (Touvron et al. 2021): Utilizing distillation to train small models by transferring knowledge from large pretrained models, reducing model complexity and computational cost.
- **MSCA** (Huang et al. 2023): A multimodal stack cross-attention network for better alignment and fusion of multimodal token-level text and visual features for the multimodal fake news detection task.
- **Multi-view CLIP** (Qin et al. 2023): A framework that is capable of leveraging multi-grained cues from multiple perspectives for the multimodal sarcasm detection task.
- **DivE** (Kim, Kim, and Kwak 2023): A cross-modal retrieval approach utilizing smoothed chamfer similarity and using ensemble prediction modules.
- **CLTL** (Wang and Markov 2024): A method that won first place for hate speech target detection in the Multimodal Hate Speech Event Detection Challenge 2024.

Evaluation Metrics

As in previous work, we also use accuracy as an evaluation metric for the model, i.e., the proportion of test samples in which the idioms selected by the model correspond to the correct idioms. Moreover, we add Avg-Acc, which represents the average accuracy of all previous test sets, to evaluate the overall performance of the model.

Results and Analysis

Comparison with Baselines

Table 3 presents the results of the comparison of our method with different baselines on the MChIRC dataset. Compared to unimodal Chinese idiom reading comprehension methods, DCIGN achieves the best accuracy on Dev, Test, Sim, and Out sets, with the highest average accuracy. In particular, our method surpasses the SOTA method by Wu et al. (2024a) in the ChID dataset. This not only indicates the challenging nature of the MChIRC dataset but also demonstrates the potential of image features to provide valuable information for the task of comprehension of Chinese idioms. Compared to the four methods used in other multimodal tasks, our method still performs optimally, demonstrating the unique applicability of our proposed DCIGN for multimodal Chinese idiom reading comprehension. At the same time, we find that the results of all methods on the Out set are less accurate than on the other test sets, which highlights the greater challenge posed by the Out set.

Method	Parameters	Acc	Team
GLM4V	9B	0.528	ZhipuAI
CogVLM2	8B	0.498	ZhipuAI
InternVL2-Pro	-	0.518	OpenGVLab
Qwen1.5	110B	0.526	Aliyun
GPT-4o	175B	0.538	OpenAI
Ours	**-**	**0.544**	-

Table 4: Comparison results of the MLLMs with 500 samples in the Out set, "B" stands for billion.

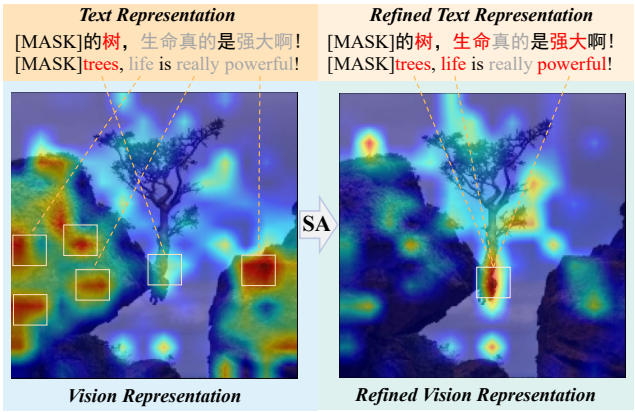
Comparison with MLLMs

We compare five advanced multimodal large language models, selected from the top 10 on the OpenCompass[6], GLM4V[7](GLM et al. 2024), CogVLM2[8], InternVL2-Pro[9], Qwen1.5[10] (Team 2024), GPT-4o[11], respectively. For cost considerations, we only test 500 data from the Out set. The comparison results are shown in Table 4. It can be observed

[6]https://rank.opencompass.org.cn/home
[7]https://github.com/THUDM/GLM-4
[8]https://github.com/THUDM/CogVLM2
[9]https://github.com/OpenGVLab/InternVL
[10]https://tongyi.aliyun.com/qianwen/
[11]https://chatgpt.com/

Method	Dev	Test	Sim	Out	Avg
w/o GCN	55.91	54.19	58.64	37.12	51.47
w/o $\mathcal{L}_{ccl}$	38.01	35.27	40.79	22.21	34.07
w/o $\mathcal{L}_{fcl}$	70.25	69.42	72.08	51.95	65.93
w/o T_{SA}	69.60	67.53	68.66	49.45	63.81
w/o V_{SA}	69.24	68.08	69.89	49.15	64.09
w/o $\mathcal{L}_{ccl}$ and $\mathcal{L}_{fcl}$	37.89	35.76	41.07	21.91	34.16
w/o T_{SA} and V_{SA}	38.52	36.22	40.98	22.11	34.46
DCIGN (Ours)	**77.26**	**77.16**	**77.34**	**60.25**	**73.00**

Table 5: Comparison of different component ablations on the MChIRC dataset. T_{SA} and V_{SA} represent the self-attention of text and image modalities, respectively.



(1) After coarse contrastive learning (2) After fine contrastive learning

Figure 5: Visual comparison presentation after contrastive contrast learning and fine contrastive learning.

that compared to MLLMs, DCIGN still exhibits strong competitiveness. This indicates that (1) there is still room for improvement in multimodal large language models' understanding of Chinese idioms in multimodal forms; (2) the model we proposed is applicable to the task of understanding multimodal Chinese idioms, providing a benchmark method for this task's first introduction.

Ablation Study

We conduct an ablation study on various components of DCIGN to validate the rationality of the network architecture. As shown in Table 5, when we individually remove $\mathcal{L}_{fcl}$, T_{SA}, and V_{SA}, the model's accuracy experienced a slight decrease. However, when we individually remove GCN, the coarse contrastive learning loss $\mathcal{L}_{ccl}$, the dual-contrastive learning loss, and the dual-modality self-attention, the model's accuracy significantly decreased. This demonstrates the effectiveness of each component of DCIGN working in concert to accomplish the task of reading comprehension for Chinese idioms.

Hyperparameter Analysis

We perform a hyperparameter analysis of the α and β before $\mathcal{L}_{ccl}$ and $\mathcal{L}_{fcl}$, and the experimental results are shown in Table 6. We find that the average accuracy of the model

α, β	Dev	Test	Sim	Out	Avg
α=0.1, β=0.9	75.50	75.32	76.11	58.13	71.27
α=0.2, β=0.8	77.40	76.16	76.74	60.15	72.61
α=0.3, β=0.7	75.26	74.60	75.48	56.87	70.55
α=0.4, β=0.6	77.26	**77.16**	**77.34**	**60.25**	**73.00**
α=0.5, β=0.5	75.60	74.28	75.33	57.96	70.79
α=0.6, β=0.4	76.49	76.04	77.01	58.58	72.12
α=0.7, β=0.3	75.79	74.01	74.96	58.37	70.78
α=0.8, β=0.2	75.48	74.39	75.48	57.48	70.71
α=0.9, β=0.1	**78.27**	76.39	76.61	60.18	72.86

Table 6: Hyperparameter analysis of α and β.

is optimized when α=0.4, β=0.6. We also find that the highest average accuracy is observed when α is higher and β is lower in the combination of α and β, e.g., α=0.9, β=0.1. On the contrary, when α is low and β is high, e.g., α=0.1, β=0.9, relatively low average accuracy is observed. This indicates that coarse contrastive learning loss accounts for a higher impact in the network optimization process, followed by fine contrastive learning loss.

Visualization

We conduct a visual comparison of text and images before and after applying coarse contrastive learning and refined contrastive learning, as shown in Figure 5. Before self-attention, coarse contrastive learning could only roughly align the text with the corresponding content in the image, as shown on the left side of Figure 5, where only the "trees" in red is aligned with the root of the tree in the image. After self-attention, fine contrastive learning allows for more precise alignment of text to image content. This refinement allows the visual presentation to focus on specific details, such as on the right side of Figure 5, where the textual analysis emphasizes the keywords "tree", "life", and "strength", which are aligned to the roots of the tree, thus effectively conveying the image of a tree thriving on the edge of a cliff. The combination of these textual and visual features reflects the indomitable vitality subtly, thus encapsulating more profoundly the essence of the idiom "Rescued from desperation". It can be concluded that we perform dual-contrastive learning on the features of text and images both before and after self-attention, thereby further enhancing the model's ability to semantically associate text-image pairs.

Conclusion

The current Chinese idiom reading comprehension methods focus mainly on text. In real-world scenarios, more multimodal forms occur simultaneously and the text length is shorter. Therefore, to address the above problems, we first construct a large-scale multimodal Chinese idiom reading comprehension dataset, MChIRC. Then, we propose a dual-contrastive idiom graph network, DCIGN. Extensive experimental results show that DCIGN can better understand the meaning of Chinese idioms with the help of their corresponding images. We believe that DCIGN, as a new benchmark for Chinese idiom reading comprehension, can provide a fresh perspective for future research.

Ethical Statement

The list of Chinese idioms used in the MChIRC dataset is from Zheng, Huang, and Sun (2019). All text-image pairs are sourced from two platforms Baidu and Sogou. We guarantee that all data are used for scientific research only. We promise that the data collected will not be used for non-commercial or profit-making purposes and strictly respect this dataset's copyright.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62272188), the Fundamental Research Funds for the Central Universities (2662021JC008), and the 2023 Huazhong Agricultural University Independent Science and Technology Innovation Fund Project (2662023XXPY005). Thanks to the anonymous reviewers for their hard efforts!

References

Ali, S.; Ravi, P.; Moore, K.; Abelson, H.; and Breazeal, C. 2024. A Picture Is Worth a Thousand Words: Co-designing Text-to-Image Generation Learning Materials for K-12 with Educators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 23260–23267.

Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2506–2515.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Cui, Y.; Liu, T.; Chen, Z.; Ma, W.; Wang, S.; and Hu, G. 2018. Dataset for the First Evaluation on Chinese Machine Reading Comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Cui, Y.; Liu, T.; Chen, Z.; Wang, S.; and Hu, G. 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1777–1786.

Dai, Y.; Liu, Y.; Yang, L.; and Fu, Y. 2023. An Idiom Reading Comprehension Model Based on Multi-Granularity Reasoning and Paraphrase Expansion. *Applied Sciences*, 13(9): 5777.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*.

Hu, Y.; Li, J.; Wu, M.; Huang, Z.; Chen, G.; and Sha, Y. 2024. Uncovering and Mitigating the Hidden Chasm: A Study on the Text-Text Domain Gap in Euphemism Identification. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 18270–18278. AAAI Press.

Huang, Z.; Hu, Y.; Zeng, Z.; Li, X.; and Sha, Y. 2023. Multimodal Stacked Cross Attention Network for Fine-Grained Fake News Detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2837–2842. IEEE.

Jiang, Z.; Zhang, B.; Huang, L.; and Ji, H. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 154–158.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23422–23431.

Knietaite, A.; Allsebrook, A.; Minkov, A.; Tomaszewski, A.; Slinko, N.; Johnson, R.; Pickard, T.; Phelps, D.; and Villavicencio, A. 2024. Is Less More? Quality, Quantity and Context in Idiom Processing with Natural Language Models. *arXiv preprint arXiv:2405.08497*.

Li, S.; Chen, J.; Yuan, S.; Wu, X.; Yang, H.; Tao, S.; and Xiao, Y. 2024. Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 18554–18563.

Long, S.; Wang, R.; Tao, K.; Zeng, J.; and Dai, X. 2020. Synonym Knowledge Enhanced Reader for Chinese Idiom Reading Comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3684–3695.

Pintado, B. R.; and Fajardo, T. 2021. Learning idioms through the multimodal approach. *Learning*, 12(24).

Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; and Chua, T.-S. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14444–14452.

Qiang, J.; Li, Y.; Zhang, C.; Li, Y.; Zhu, Y.; Yuan, Y.; and Wu, X. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11: 740–754.

Qin, L.; Huang, S.; Chen, Q.; Cai, C.; Zhang, Y.; Liang, B.; Che, W.; and Xu, R. 2023. MMSD2. 0: Towards a Reliable Multi-modal Sarcasm Detection System. In *Findings of the Association for Computational Linguistics: ACL 2023*, 10834–10845.

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

Sha, Y.; Wu, M.; Zeng, Z.; Ge, X.; Huang, Z.; and Wang, H. 2023. A Prompt-Based Representation Individual Enhancement Method for Chinese Idiom Reading Comprehension. In *International Conference on Database Systems for Advanced Applications*, 682–698. Springer.

Shao, Y.; Sennrich, R.; Webber, B.; and Fancellu, F. 2018. Evaluating Machine Translation Performance on Chinese Idioms with a Blacklist Method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

Su, G.; Wu, M.; Huang, Z.; Zhang, Y.; Wang, T.; Hu, Y.; and Sha, Y. 2024. Refine, Align, and Aggregate: Multi-view Linguistic Features Enhancement for Aspect Sentiment Triplet Extraction. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 3212–3228. Association for Computational Linguistics.

Tan, M.; and Jiang, J. 2021. Learning and evaluating Chinese idiom embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1387–1396.

Team, Q. 2024. Introducing Qwen1.5.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, X.; Li, C.; Zhao, J.; and Yu, D. 2021. NaturalConv: A Chinese Dialogue Dataset Towards Multi-turn Topic-driven Conversation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 14006–14014. AAAI Press.

Wang, X.; Zhao, H.; Yang, T.; and Wang, H. 2020. Correcting the misuse: A method for the Chinese idiom cloze test. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 1–10.

Wang, Y.; and Markov, I. 2024. CLTL@ Multimodal Hate Speech Event Detection 2024: The Winning Approach to Detecting Multimodal Hate Speech and Its Targets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, 73–78.

Wong, L. H.; Chin, C. K.; Tan, C. L.; Liu, M.; and Gong, C. 2010. Students' meaning making in a mobile assisted Chinese idiom learning environment. In *Proceedings of the 9th International Conference of the Learning Sciences-Volume 1*, 349–356.

Wu, M.; Hu, Y.; Zhang, Y.; Zhi, Z.; Su, G.; and Sha, Y. 2024a. Mitigating Idiom Inconsistency: A Multi-Semantic Contrastive Learning Method for Chinese Idiom Reading Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 19243–19251.

Wu, M.; Su, G.; Zhang, Y.; Huang, Z.; and Sha, Y. 2024b. Refining Idioms Semantics Comprehension via Contrastive Learning and Cross-Attention. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13785–13795.

Yosef, R.; Bitton, Y.; and Shahaf, D. 2023. IRFL: Image Recognition of Figurative Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1044–1058.

Zhang, Q. 2021. A Study on the Application of Multimodal Metaphor in English Idiom Teaching. In *2nd International Conference on Education Studies: Experience and Innovation (ICESEI 2021)*, 111–115. Atlantis Press.

Zhang, Y.; Kong, L.; Tian, S.; Fei, H.; Xiang, C.; Wang, H.; and Wei, X. 2024. Multi-view Counterfactual Contrastive Learning for Fact-checking Fake News Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, June 10-14, 2024*, 385–393. ACM.

Zheng, C.; Huang, M.; and Sun, A. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 778–787.