Enhancing Chinese Multimodal Entity Linking with CLIP-RoBERTa and Contrastive Learning

Chun Wang[1], Chunyan An[1] (✉), Qiang Yang[2], and Zhixu Li[3,4]

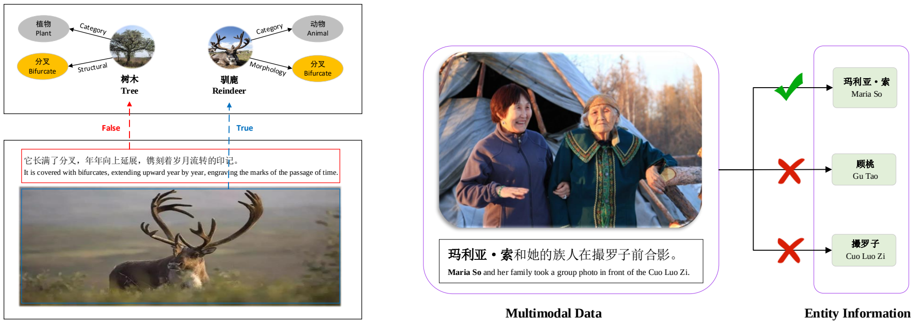
[1] School of Computer Science, Inner Mongolia University, Hohhot, China
chunwang@mail.imu.edu.cn, ann@imu.edu.cn
[2] University of Florida, Gainesville, United States
qiangyang@ufl.edu
[3] School of Information, Renmin University of China
[4] School of Smart Governance, Renmin University of China
zhixuli@ruc.edu.cn

Abstract. Multimodal Entity Linking (MEL) is a vital task in natural language processing, aiming to associate ambiguous mentions in multimodal datasuch as text and images with entities in a knowledge base (KB). However, existing methods focus primarily on English corpora, limiting their applicability to Chinese data. This challenge is further exacerbated by the scarcity of Chinese text-image datasets and semantic inconsistencies between English and Chinese languages, making it difficult to adapt English-based models for Chinese contexts. Short text labels often lack semantic richness, while noise in visual data introduces irrelevant features, reducing linking accuracy. To address these challenges, we propose MCR, a contrastive learning-based Multimodal Entity Linking method tailored for Chinese datasets. MCR leverages CLIP-RoBERTa for deep feature learning and incorporates contrastive learning to strengthen feature relationships, enabling more accurate linking of multimodal data. Additionally, we contribute a high-quality Chinese MEL dataset focused on ethnic minority elements, addressing a critical resource gap. Experimental results on this new dataset and other benchmarks demonstrate that MCR outperforms existing methods across multiple metrics, establishing it as a robust solution for MEL in both Chinese and English domains.

Keywords: Multimodal Entity Linking · Knowledge Graph · Chinese Context · Contrastive Learning.

1 Introduction

Entity Linking (EL) is the task of aligning ambiguous mentions in text to their corresponding entities in knowledge graphs (KGs) [1], serving as a crucial bridge between unstructured content and structured knowledge. EL plays a vital role in numerous information retrieval and natural language processing applications, including question answering [2], semantic search [3, 4], content analysis [5, 6], information extraction [7] and recommendation systems [8]. The ambiguity of mentions may pose a great challenge to EL due to multiple candidate entities, insufficient or noisy contextual information. Traditional Entity Linking (EL) methods [9–11] cannot adequately resolve these ambiguities, as they often rely heavily on limited textual context and struggle to handle cases with incomplete or inconsistent information. Fortunately, with the rapid development and widespread use of social media and the Internet, textual and visual multimodality has become a crucial medium for data-driven tasks. For example, as shown in Figure 1 (a), without the provided image, the mention will be linked to the entity "Tree" since it shares the very similar semantics from the context text. But with the help of the information from the image e.g., "Reindeer", the mention will be correctly linked to the Reindeer existing in the knowledge base. Therefore, Multimodal Entity Linking (MEL) is introduced, which aims to link mentions using both textual and visual context to their corresponding entities in multimodal knowledge bases [12–14]. However, the quality of online information may not be consistent where many mentions are ambiguous and their context are not complete. This exacerbates the difficulty of accurately disambiguating mentions, underscoring the need for advanced methods that integrate textual and visual modalities to improve precision in EL.



它长满了分叉，年年向上延展，镌刻着岁月流转的印记。
It is covered with bifurcates, extending upward year by year, engraving the marks of the passage of time.

玛利亚・索和她的族人在撮罗子前合影。
Maria So and her family took a group photo in front of the Cuo Luo Zi.

(a) An example of multimodal entity linking.

(b) An example of the influence of short text labels in MEL.

Fig. 1: Running Examples.

Numerous deep learning-based methods have been introduced to integrate visual information with textual context for linking multimodal mentions to entities [15–18]. For eaxmple, Adjali et al. proposed to jointly learn a representation of both mentions and entities from their textual and visual contexts [15]. Zhang et al. put forward a novel MEL model that effectively mitigated the negative impact of noisy images while leveraging multiple attention mechanisms to enhance the connection between mention representations and their corresponding entity representations [17]. Although these methods have achieved notable progress in MEL tasks, they still face several critical challenges:

- The scarcity of Chinese text-image datasets and semantic inconsistencies hinder the adaptation of English-trained models to Chinese contexts.
- The lack of semantic richness in short text labels often provide insufficient context for disambiguating entities with similar characteristics.
- In the visual modality, images often contain a mix of target entities or irrelevant visual information, introducing noise that complicates accurate linking. For instance, short text labels like "Maria So" and the presence of visual noise in Figure 1 (b) make it difficult to identify the intended entity.

Some methods or datasets are proposed to solve these mentioned challenges. For example, Zhou et al. constructed a mixed MEL dataset containing Weibo-MEL (for Chinese), Wikidata-MEL (for English) and Richpedia-MEL (for English) collected from social media, encyclopedia and multimodal knowledge graphs respectively. To address the second challenge, Cheng et al. [19] proposed the BERT-ENE model, which extracted descriptive texts of entities from the knowledge base to generate rich vector embeddings. For the third one, Li et al. [20] proposed UniFormer, which unified convolution and self-attention in a concise transformer architecture to effectively mitigate visual redundancy and dependency. Although these methods address some challenges to a certain extent, they cannot effectively handle the integration of textual and visual information, particularly in scenarios involving noisy data, ambiguous short text labels, or incomplete knowledge bases. Fortunately, CLIP model [21], trained on English image-text datasets, can align visual and textual features in a shared embedding space. However, it cannot be directly applied to Chinese scenarios due to significant differences in language structure and data distribution. To bridge this gap, Yang et al. proposed the Chinese-CLIP method [22], which leveraged the visual parameters of the English CLIP model and the language parameters of the Chinese RoBERTa model as initialization. Moreover, these approaches may lack robust mechanisms for aligning modalities in a unified representation space. In addition, the quality of existing Chinese MEL data further restricts the training and evaluation of these models.

To address these challenges, we propose MCR (Multimodal Entity Linking with CLIP-RoBERTa), a contrastive learning-based framework designed to enhance multimodal representation capabilities for the Chinese context. By leveraging the feature extraction strengths of CLIP-RoBERTa and employing contrastive learning, MCR effectively bridges the gap between textual and visual modalities, tackling the issues of noisy data and ambiguous short text labels. Moreover, the construction of a high-quality, diverse Chinese multimodal dataset provides a strong foundation for training and evaluation, paving the way for more accurate and robust entity linking in the Chinese domain. Extensive experiments, evaluated on a newly constructed dataset, demonstrate that MCR outperforms existing methods across multiple metrics, establishing it as a robust solution for MEL in both Chinese and English domains.

Our contributions can be summarized as follows:

- We propose a contrastive learning-based MEL method, MCR, which effectively integrates information from text and image modalities, enhancing representation capabilities and addressing gaps in existing Chinese MEL research.

- We contribute a newly constructed Chinese multimodal entity linking dataset, called EMMEL, a high-quality Chinese multimodal data, accounting for 10,028 pairs.
- Experimental results on a newly constructed Chinese multimodal entity linking dataset, EMMEL, demonstrate that MCR significantly outperforms existing methods across multiple metrics. Moreover, the proposed method achieves robust performance improvements in both Chinese and English domains, establishing it as a versatile solution for MEL.

2 Related Work

2.1 Multimodal Entity Linking Methods

Early work on MEL was initiated by Moon et al. [23], who proposed a zero-shot framework that integrates textual, visual, and lexical information to link entities in social media posts. Building on this, Adjali et al. introduced a model that jointly learns mention and entity representations from both textual and visual contexts, incorporating statistical information for enhanced linking [15]. Zhang et al. developed a two-stage mechanism to filter irrelevant images and employed attention mechanisms to refine mention and entity representations using multi-hop connections [17]. Cui et al. proposed a Transformer-based model that estimates entity-mention connections by learning joint image-text representations in a shared embedding space [24]. Wang et al. introduced Gated Hierarchical Multimodal Fusion with Contrastive Training (GHMFC), which utilizes co-attention mechanisms to discover fine-grained cross-modal correlations [12]. Zheng et al. employed a multimodal feature engineering approach, integrating visual-textual features from both corpora and knowledge graphs, and achieved state-of-the-art results with a two-phase training strategy [25].

Recent advancements include Wang et al. [26], who introduced Diverse-Modal Entity Linking (DMEL), addressing text, image, and table modalities with a generative encoder-decoder framework. Shi et al. proposed GEMEL, a framework based on large language models (LLMs) that directly generates entity names [14]. Li et al. developed a multimodal fusion module with contrastive learning to capture bilinear interactions [27], while Xing et al. proposed a dynamic relationship interaction framework based on Graph Convolutional Networks for more accurate feature alignment [28].

Despite these advancements, most methods focus on English datasets and fail to address challenges specific to the Chinese domain, such as noisy visual data, ambiguous short text labels, and limited cross-modal integration. Additionally, their reliance on large-scale datasets and computationally intensive training pipelines restricts their adaptability.

2.2 Multimodal Entity Linking Datasets

Adjali et al. released an annotated dataset of Twitter posts and proposed a model for jointly learning a representation of both mentions and entities from their textual and visual contexts [15]. Gan et al. introduced a dataset based on movie reviews, primarily focusing on character mentions in the film domain [29]. Zhou et al. constructed a mixed MEL dataset containing Weibo-MEL (for Chinese), Wikidata-MEL (for English) and Richpedia-MEL (for English) collected from social media, encyclopedia and multimodal knowledge graphs [30]. Wang et al. proposed WikiDiverse, a manually annotated dataset that covers diverse topics and entity types [18]. More recently, Yao et al. developed AMELI, a large-scale dataset consisting of 18,472 reviews and 35,598 products [31].

While these datasets contribute to the MEL field, there is a notable shortage of high-quality Chinese multimodal datasets. This lack of resources limits the development and evaluation of MEL models tailored to the Chinese domain, underscoring the need for curated datasets with diverse and representative multimodal information.

3 EMMEL Dataset Construction

The EMMEL (Ethnic Minority Multimodal Entity Linking Dataset) is a Chinese multimodal entity linking dataset built through a unified annotation system, focusing on the field of ethnic minority culture, and containing rich image and text information.

3.1 Data Source Selection

Given the scarcity of resources in this domain, we carefully selected diverse data sources to ensure a rich and representative dataset. First, we scanned the book “Evenk Traditional Society and Culture” [32], curated relevant content, and annotated it to emphasize ethnic minority elements. However, relying on data from a single ethnic group could lead to limited representativeness. To address this, we supplemented the dataset by scraping content from intangible cultural heritage websites in regions such as Guangdong and Guizhou. Additionally, recognizing that ethnic minority cultures encompass not only traditions and customs but also diverse fauna, we refined an existing English dataset to enhance the animal-related elements, enriching the overall dataset.

3.2 Data Acquisition

We utilized image-title pairs with accompanying descriptions covering diverse themes such as folklore, production, branches, history, scenic spots, religious beliefs, and intangible cultural heritage projects, ensuring comprehensive coverage of common topics in the ethnic minority domain. Additionally, we manually refined the ontology and entity relationships of existing textual knowledge graphs in this domain to enhance their structure and relevance, providing robust support for multimodal entity linking tasks.

3.3 Data Cleaning and Augmentation

For image-title pairs, we removed entries with low-resolution images or mismatched content to ensure data quality. Additionally, we applied data augmentation techniques to the images, including specific methods such as rotation and translation, to enhance the diversity and robustness of the dataset.

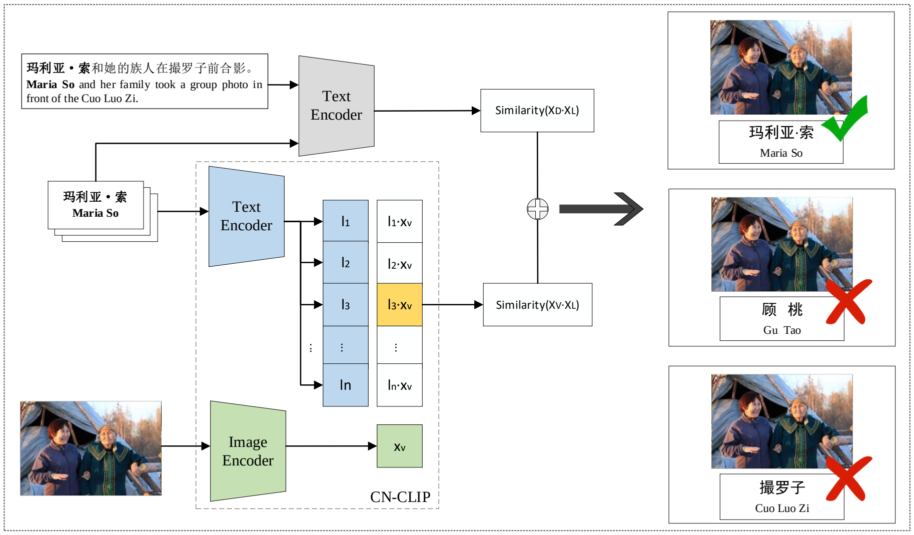


Fig. 2: Overview of the proposed MCR method.

4 Methodology

The overall framework of the proposed model is illustrated in Figure 2, which demonstrates the process of linking multimodal data to corresponding entities in a KB. Given a multimodal input comprising an image and its descriptive text, the model begins by extracting contextual features from the text using RoBERTa and visual features from the image using CN-CLIP (Section 4.2). Subsequently, the framework calculates two types of similarities: the similarity between the descriptive text and entity labels in the knowledge graph (Section 4.3), and the similarity between the image and entity labels (Section 4.4). Finally, these similarity features are fused to predict the most relevant entity in the knowledge graph (Section 4.5), ensuring robust and accurate multimodal entity linking.

4.1 Problem Definition

Let $X = \{x_i\}_{i=1}^{N}$ be a set of N input multimodal samples,with corresponding entities set $Y = \{e_i\}_{i=1}^{M}$ in a KG. Each input sample is composed of three parts: $x = \{x_v; x_d; x_l\}$, where x_v represents the image, x_d is the description text associated with the image, and x_l represents the textual entity label e_i.

The goal of the Multimodal Entity Linking (MEL) task is to link ambiguous mentions to entities in the KG by computing the similarity between joint multimodal mention features and the textual representations of entities in the KG. The entity with the highest similarity score is selected as the corresponding linked entity. This process leverages both textual and visual information to resolve ambiguities and clarify semantic meanings. Finally, The most relevant KG entity y is selected from the λ candidate entities $C_e = \{e_i\}_{i=1}^{\lambda}$. This can be expressed as follows:

$$y = \underset{\forall e \in C_e}{\arg\max}\, \mathit{\Gamma}\left(\mathit{\Phi}(x_v, x_l), \mathit{\Psi}(x_d, x_l)\right) \tag{1}$$

where $\mathit{\Phi}$ represents the feature function of the multimodal encoder module, $\mathit{\Psi}$ represents the feature function of the text encoder module, and $\mathit{\Gamma}$ represents the similarity score between the multimodal and text entities. y is the predicted linked entity.

4.2 Textual and Visual Contexts Features

Textual features. Given an entity x_l and its description x_d, we tokenize them into a sequence of word embeddings using RoBERTa [33]. In this sequence, the special tokens $[CLS]$ and $[SEP]$ are used to mark the beginning of the sequence and the separator, respectively. The generated word embedding sequence is then fed into RoBERTa's Transformer encoder, which models the contextual semantics of the input. Through its multi-layer attention mechanism, the encoder outputs high-dimensional contextual representations for each subword.

Visual features. Given an entity x_l and an image x_i, we follow the contrastive learning method used in CN-CLIP [22] for training. The image encoder from CN-CLIP is used to extract visual features, encoding both image entities and textual entities into vector representations. The similarity between each image vector and its corresponding text vector is computed using the dot product, enabling a robust alignment of visual and textual features.

4.3 Text-Text Embeddings

This module leverages the pre-trained RoBERTa model, using cross-validation on the small-sample dataset to ensure model stability and generalization. To emphasize the semantics of the target text during training, the candidate entity set from the Knowledge Graph (KG) is concatenated with the target text to form a single input sentence for the model. Firstly, the entity labels and the image description text are concatenated as a sentence pair, separated by $[SEP]$. Special tokens $[CLS]$ and $[SEP]$ are added at the beginning and end of the sequence, respectively. The input format is represented as follows: $input =$

$\{[CLS], l_1, \ldots, l_i, \ldots, l_n, [SEP], d_1, \ldots, d_m, [SEP]\}$ where $X_l = \{l_1, \ldots, l_n\}$ represent the entity labels in the KG, and $X_d = \{d_1, \ldots\ d_m\}$ represent the image description text. The input represents the concatenated sequence. The $[CLS]$ token is used to determine whether the short text and entity description text are in the same semantic space. The first $[SEP]$ is used to concatenate the two sentences and distinguish the sentence pair, while the second $[SEP]$ marks the end of the input text. Next, the word vectors corresponding to the input are fed into the encoding layer of the RoBERTa model, and the feature vector corresponding to the $[CLS]$ token in the final hidden state of the RoBERTa model is extracted. The feature vector format is represented as follows:

$$H_{\text{input}} = \text{Transformer}(\text{input}) \tag{2}$$

$$H_{[\text{CLS}]} = \text{retain}(H_{\text{input}}) \tag{3}$$

where $H_{\text{input}} \in R^{l \times d}$ represents the final hidden state of the pre-trained language model, where l is the length of the input and d is the vector dimension output by Transformer. The function retain$(\cdot)$ denotes retaining the feature vector corresponding to the first $[CLS]$ token. Finally, $H_{[\text{CLS}]}$ is fed into a fully connected linear layer to compute the correlation score between the image description text x_d and the candidate entity label x_l, as shown below:

$$S_{x_l, x_d} = H_{[\text{CLS}]} W^T + b \tag{4}$$

where W is the weight matrix, and b is the bias vector.

4.4 Visual-Text Embeddings

This module comprises two independent encoders and adopts the contrastive learning approach from CN-CLIP for training. By jointly learning features from both visual and textual modalities, the module generates robust cross-modal representations. These representations enhance the model's ability to effectively align and integrate multimodal information, providing stronger representational capabilities for the entity linking task. Firstly, the textual modality extracts semantic features through the text encoder RoBERTa. Given the input entity label x_l, it is processed by the encoder f_l to generate the text embedding E_l:

$$E_l = f_l(x_l) \tag{5}$$

where E_l is the high-dimensional vector representation of the text, capturing contextual semantics and linguistic features. Next, the visual modality extracts visual features through the pre-trained visual encoder CN-CLIP. Given the input image x_v, it is processed by the encoder f_v to generate the visual embedding E_v:

$$E_v = f_v(x_v) \tag{6}$$

where E_v is the high-dimensional vector representation of the image. Then, the joint embedding is optimized using the InfoNCE contrastive loss function, bringing modal features related to the target entity closer together while pushing unrelated modal features further apart:

$$L = -\log \frac{\exp(E_v \cdot E_{l+}/\tau)}{\sum_{i=0}^{N} \exp(E_v \cdot E_{li}/\tau)} \tag{7}$$

where E_v and E_{l+} represent the feature vector from the image encoder f_v and the positive feature vector from the text encoder f_l, corresponding to the matching visual feature. E_{li} is a set of textual features including both the positive and negative features (non-matching pairs). τ is a temperature hyperparameter that controls the sharpness of the similarity distribution. The numerator denotes the similarity between positive examples, while the denominator represents the similarity between positive and negative examples. Further, the input image x_v and entity label x_l are encoded by the previously trained Encoder and Decoder. Finally, the similarity between the image features and text features is calculated, yielding the relevance score:

$$S_{x_l,x_v} = sim(E_l, E_v) \tag{8}$$

where sim(·) denotes the similarity function.

4.5 Similarity Fusion

In MEL tasks, similarity scores derived from image and text modalities can be fused to fully leverage the complementary strengths of both modalities. To achieve this, the similarity between image features and text features is combined using a weighted average, resulting in a final comprehensive similarity score:

$$y = \alpha \cdot S_{x_l,x_v} + (1-\alpha) \cdot S_{x_l,x_d} \tag{9}$$

where α is a hyperparameter that balances the contributions of the two modalities, S_{x_l,x_v} represents the similarity from the image modality, and S_{x_l,x_d} represents the similarity from the text modality.

5 Experiments

In this section, we first introduce the used datasets, and then explain the comparative baseline methods, followed the evaluation metrics. The experiments setup is introduced for the reproduction of our model. Finally, we report the experimental results of our MCR and other baselines.

5.1 Datasets

To evaluate the performance of MEL models comprehensively, we utilized multiple datasets spanning both Chinese and English domains. In addition to our constructed Chinese MEL dataset, EMMEL, we incorporated the COCO-CN dataset [34], an enhanced version of MS-COCO. This dataset includes manually written Chinese captions and labels that align visual content with textual descriptions, offering a valuable resource for multimodal tasks in the Chinese context. To further validate the robustness of MEL models across different

Table 1: The statistics of multimodal datasets

Chinese Datasets				**English Datasets**			
Dataset	**Sample**	**Entity**	**Text Len.**	**Dataset**	**Sample**	**Entity**	**Text Len.**
EMMEL	10,028	751	16.1	Wikidiverse	7,025	5,058	10.2
COCO-CN	20,341	1,932	16.8	WikiMEL	20,900	15,033	8.2

languages and domains, we included two widely used English MEL datasets: WikiDiverse [18] and WikiMEL [35]. WikiDiverse is a high-quality, manually annotated dataset covering a wide range of topics, entity types, and ambiguities. WikiMEL, on the other hand, leverages the extensive knowledge base of Wikidata, providing a challenging benchmark for linking mentions to entities with rich contextual information. The statistical details of these datasets are summarized in Table 1.

5.2 Baselines

For Chinese MEL task, we compared our model MCR with CN-CLIP with the versions of zero-shot and fine-tuning settings. CN-CLIP is a vision-language foundation model tailored for Chinese, pretrained on a 200-million-sample dataset using a two-stage method to improve efficiency and effectiveness [36]. It trained the model with locked-image tuning in the first stage and contrastive tuning in the second one. To further evaluate the effectiveness of our proposed method, we adapted MCR to an English model and conducted comparative experiments with several existing English multimodal entity linking models. JMEL utilizes fully connected layers to project both visual and textual features into latent spaces for multimodal alignment [15]. ARNN employs an AttentionRNN to predict the association between the textual features of candidate entities and the input features, focusing on enhancing semantic relevance [37]. MEL-HI incorporates multi-head attention mechanisms to capture richer contextual information and mitigate the negative impact of noisy images [17]. DWE leverages fine-grained image attributes, such as facial features and scene characteristics, to enhance and refine visual feature representations for better multimodal alignment [38]. GHMFC employs gated multi-channel fusion and a novel attention mechanism to effectively integrate multi-channel entity representations for precise linking [12]. MMEL features a joint feature extraction module to learn representations of context and entity candidates from both visual and textual modalities, enhancing multimodal entity linking performance [39]. All these methods are evaluated using accuracy metrics, specifically Top-1, Top-5, and Top-10 accuracy (%).

5.3 Experiments Setup

During the experiments, the dimensions of the text representations x_t, x_d, and the visual representation x_v were set to 512. The dimension of the relevant visual object features R was set to 768, and the number of heads in the Multihead Attention was set to 8.Other specific hyperparameters are shown in Table 2.

Table 2: Hyper-parameter Settings

Dataset	**lr**	**wd**	**batch size**	α	**vision model**
EMMEL	1e-5	0.001	64	0.5	ViT-B/16
COCO-CN	1e-5	0.001	32	0.4	ViT-L/14
Wikidiverse	1e-5	0.001	64	0.5	ViT-B/16
WikiMEL	1e-5	0.001	64	0.5	ViT-B/16

5.4 Results

Experimental results on Chinese MEL datasets. Table 3 presents the experimental results of the evaluated models, including CN-CLIP (zero-shot), CN-CLIP (fine-tuned), and the proposed MCR, on the EMMEL and COCO-CN datasets. The evaluation metrics, Top-1, Top-5, and Top-10 accuracy (%), measure the proportion of correct entities ranked within the top 1, 5, or 10 predictions out of a candidate set (λ=100).

The results demonstrate that MCR consistently outperforms both versions of CN-CLIP across all metrics and datasets. On the EMMEL dataset, MCR achieves a Top-1 accuracy of 77.8%, significantly surpassing CN-CLIP (fine-tuned) at 40.4% and CN-CLIP (zero-shot) at 28.5%. It also achieves near-perfect Top-5 accuracy (96.4%) and perfect Top-10 accuracy (100.0%), showcasing its robustness in linking entities accurately. Similarly, on the COCO-CN dataset, MCR delivers the best performance with a Top-1 accuracy of 71.7%, compared to 34.2% for CN-CLIP (fine-tuned) and 21.8% for CN-CLIP (zero-shot). MCR also excels in Top-5 and Top-10 accuracy, achieving 87.1% and 90.4%, respectively, outperforming CN-CLIP in both cases. These results highlight MCR's ability to effectively leverage multimodal features, overcome noisy and ambiguous data, and achieve superior performance in Chinese multimodal entity linking tasks.

Table 3: MEL performance on the Chinese MEL datasets at Top-1, 5, 10 accuracies (%). The best results are highlighted in bold.

Model	**EMMEL**			**COCO-CN**		
	Top-1	**Top-5**	**Top-10**	**Top-1**	**Top-5**	**Top-10**
CN-CLIP$_{\text{zero-shot}}$	28.5	55.7	66.8	21.8	50.8	65.7
CN-CLIP$_{\text{fine-tuning}}$	40.4	77.5	82.6	34.2	71.7	85.9
MCR	**77.8**	**96.4**	**100.0**	**71.7**	**87.1**	**90.4**

Experimental results on English MEL datasets. Table 4 presents the performance of MEL models on two English MEL datasets, WikiDiverse and WikiMEL, evaluated using Top-1, Top-5, and Top-10 accuracy metrics. Across both datasets, the proposed MCR consistently outperforms other baseline methods, demonstrating its effectiveness in entity linking tasks. On WikiDiverse, MCR achieves the highest Top-1, Top-5, and Top-10 accuracies at 58.4%, 97.1%,

Table 4: MEL performance on the English MEL datasets at Top-1, 5, 10 accuracies (%). The best results are highlighted in bold. '-' denotes the results are not available.

Model	**Wikidiverse**			**WikiMEL**		
	Top-1	**Top-5**	**Top-10**	**Top-1**	**Top-5**	**Top-10**
JMEL	21.9	54.5	69.9	31.3	49.4	57.9
ARNN	22.4	50.5	68.4	32.0	45.8	56.6
MEL-HI	27.1	60.7	78.7	38.7	55.1	65.2
GHMFC	-	-	-	43.6	64.0	74.4
MMEL	-	-	-	71.5	91.7	96.3
DWE	51.2	91.0	96.3	72.8	**97.5**	**98.9**
MCR	**58.4**	**97.1**	**97.9**	**74.6**	83.8	86.2

and 97.9%, respectively, outperforming the strongest baseline, DWE, which achieves 51.2%, 91.0%, and 96.3% on the same metrics. Similarly, on WikiMEL, MCR delivers the best Top-1 accuracy (74.6%) and competitive Top-5 and Top-10 accuracies (83.8% and 86.2%), while DWE achieves slightly higher Top-5 (97.5%) and Top-10 (98.9%) results.

The superior performance of MCR can be attributed to its advanced contrastive learning mechanism and the effective integration of multimodal features, which enhance its ability to resolve ambiguities and leverage complementary information from textual and visual data. While methods like DWE and MEL-HI focus on refining visual attributes, MCRs joint learning of textual and visual modalities ensures robust cross-modal representations. The slightly lower Top-5 and Top-10 accuracy on WikiMEL compared to DWE could be due to DWE's specific emphasis on fine-grained visual features with detailed image attributes.

5.5 Ablation Study

To evaluate the contribution of each module to the performance of the MCR model in the multimodal entity linking task, we conducted ablation experiments by sequentially removing core components and analyzing their impact on the model's accuracy. As shown in Figure 3, the results clearly illustrate the significance of each module in the overall performance of MCR. When the Text-Text Module is removed, the model experiences a substantial decline in performance, with Top-1 accuracy dropping from 77.8% to 63.4%, Top-5 accuracy decreasing from 96.4% to 79.6%, and Top-10 accuracy reducing from 100% to 85.3%. This highlights the pivotal role of the Text-Text Module in capturing textual semantics and aligning mentions with entities in the knowledge graph. In contrast, removing the Visual-Text Module results in a more gradual decline in accuracy, with Top-1 accuracy falling to 70.3%, Top-5 accuracy to 93.4%, and Top-10 accuracy to 97.6%. Although the Visual-Text Module also contributes significantly to the performance by aligning visual information with text, its impact is less pronounced compared to the Text-Text Module. These results demonstrate

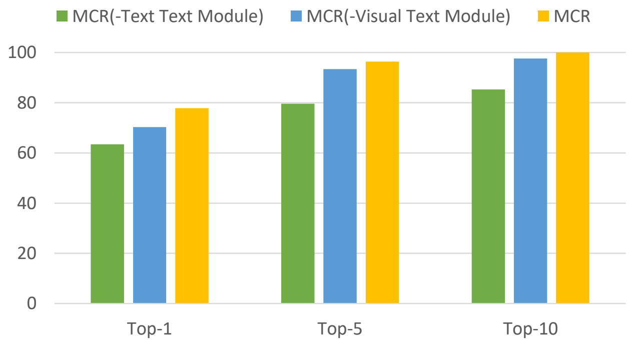


Fig. 3: Ablation analysis of MCR.

that while both modules are critical for the multimodal entity linking task, the Text-Text Module plays a more dominant role in improving performance.

5.6 Case Study

To evaluate the effectiveness of the MCR model, we conducted a case study based on examples shown in Figure 4, analyzing both correct and incorrect predictions. For the correct predictions, Case 1 highlights the model's ability to handle images with multiple visual objects, including noisy data, by successfully identifying the primary element (a person) through visual features and accurately linking it to the entity "Maria So" with the support of textual information. Similarly, in Cases 3 and 4, the model distinguishes between residential spaces with similar functionalities, demonstrating its capacity to integrate textual and visual contexts effectively for precise entity identification. For the incorrect predictions, Case 2 reveals the model's potential despite the error. The input included multiple elements, such as “Evenk wedding” and “traditional ethnic clothing”, resulting in high similarity scores for both. However, the similarity fusion mechanism failed to effectively determine which features should dominate the prediction, leading to an incorrect outcome. This demonstrates the model's ability to learn meaningful relationships between textual and visual features, even when the final prediction is incorrect. These cases illustrate that MCR effectively integrates multimodal information, leveraging similarity fusion to refine entity identification.

6 Conclusion

In this work, we proposed MCR, a contrastive learning-based multimodal entity linking method tailored to address the challenges of linking short-text entities in the Chinese domain. To support this task, we constructed the EMMEL dataset through an in-depth analysis and curation of ethnic minority-related data, filling a critical gap in existing Chinese multimodal entity linking resources. Experimental results demonstrate that MCR excels in integrating textual and visual

Case	1	2	3	4
Image				
Text	**玛利亚・索**和她的族人在撮罗子前合影 **Maria So** and her family took a group photo in front of the Cuo Luo Zi	**鄂温克族婚礼**中，新娘与伴娘合影 A photo of the bride and bridesmaids at an **Evenk wedding**	**鄂温克包**，鄂温克人居住的包 **Evenk yurt**, the traditional dwelling of the Evenk people	**撮罗子**，外形如圆锥形的结构建筑 **Cuo Luo Zi** is a conical-shaped structural building
Entity	**玛利亚・索** **Maria So**	**民族服饰** **Traditional ethnic clothing**	**鄂温克包** **Evenk yurt**	**撮罗子** **Cuo Luo Zi**
Predict	✓	×	✓	✓

Fig. 4: Running cases of our MCR method.

information, achieving superior performance on Chinese datasets while also delivering notable improvements across several metrics in the English domain.

Acknowledgments. The authors wish to thank the Projects of Natural Science Foundation of Inner Mongolia under Grant No. 2024MS06014, Inner Mongolia Science & Technology Plan under Grant No. 2021GG0164, Natural Science Foundation of China under Grant No. 62162046, 62362054, 62366038, Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), the Engineering Research Center of Ecological Big Data, Ministry of Education, China.

References

1. Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, et al. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*, 2024.
2. Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*, 2021.
3. Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 179–188, 2015.
4. Edgar Meij, Krisztian Balog, and Daan Odijk. Entity linking and retrieval for semantic search. *WSDM*, 10:2556195–2556201, 2014.
5. Mauricio Marrone, Sascha Lemke, and Lutz M Kolbe. Entity linking systems for literature reviews. *Scientometrics*, 127(7):3857–3878, 2022.
6. Qiang Yang, Xiaodong Wu, Xiuying Chen, Xin Gao, and Xiangliang Zhang. Think as people: Context-driven multi-image news captioning with adaptive dual attention. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4430–4434, 2024.
7. Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web: a survey. *Semantic Web*, 11(2):255–335, 2020.
8. Defu Lian, Qi Liu, and Enhong Chen. Personalized ranking with importance sampling. In *Proceedings of The Web Conference 2020*, pages 1093–1103, 2020.

9. Feng Wei, Uyen Trang Nguyen, and Hui Jiang. Dual-fofe-net neural models for entity linking with pagerank. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28*, pages 635–645. Springer, 2019.
10. Chen Liu, Feng Li, Xian Sun, and Hongzhe Han. Attention-based joint entity linking with entity embedding. *Information*, 10(2):46, 2019.
11. Wei Shi, Siyuan Zhang, Zhiwei Zhang, Hong Cheng, and Jeffrey Xu Yu. Joint embedding in named entity linking on sentence level. *arXiv preprint arXiv:2002.04936*, 2020.
12. Peng Wang, Jiangheng Wu, and Xiaohang Chen. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 938–948, 2022.
13. Qi Liu, Yongyi He, Tong Xu, Defu Lian, Che Liu, Zhi Zheng, and Enhong Chen. Unimel: A unified framework for multimodal entity linking with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1909–1919, 2024.
14. Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. Generative multimodal entity linking. *arXiv preprint arXiv:2306.12725*, 2023.
15. Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Multimodal entity linking for tweets. In *European Conference on Information Retrieval*, pages 463–478. Springer, 2020.
16. Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1583–1594, 2023.
17. Li Zhang, Zhixu Li, and Qiang Yang. Attention-based multimodal entity linking with high-quality images. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26*, pages 533–548. Springer, 2021.
18. Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. Wikidiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. *arXiv preprint arXiv:2204.06347*, 2022.
19. J Cheng, C Pan, J Dang, Z Yang, X Guo, L Zhang, and F Zhang. Entity linking for chinese short texts based on bert and entity name embeddings. In *China Conference on Knowledge Graph and Semantic Computing (CCKS)*, 2019.
20. Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023.
21. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
22. An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.
23. Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity disambiguation for noisy social media posts. In *Annual Meeting of the Association for Computational Linguistics*, pages 2000–2008, 2018.

24. Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021.
25. Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. Visual entity linking via multi-modal learning. *Data Intelligence*, 4(1):1–19, 2022.
26. Sijia Wang, Alexander Hanbo Li, Henry Zhu, Sheng Zhang, Chung-Wei Hang, Pramuditha Perera, Jie Ma, William Wang, Zhiguo Wang, Vittorio Castelli, et al. Benchmarking diverse-modal entity linking with generative models. *arXiv preprint arXiv:2305.17337*, 2023.
27. Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. Imf: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, pages 2572–2580, 2023.
28. Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3599–3608, 2023.
29. Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. Multimodal entity linking: a new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 993–1001, 2021.
30. Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. Weibo-mel, wikidata-mel and richpedia-mel: multimodal entity linking benchmark datasets. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 315–320. Springer, 2021.
31. Barry Menglong Yao, Yu Chen, Qifan Wang, Sijia Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. Ameli: Enhancing multimodal entity linking with fine-grained attributes. *arXiv preprint arXiv:2305.14725*, 2023.
32. Limin Bai. *Evenk Traditional Society and Culture.* 2007.
33. Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.
34. Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.
35. Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.
36. An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.
37. Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. Named entity disambiguation for noisy text. *arXiv preprint arXiv:1706.09147*, 2017.
38. Shezheng Song, Shan Zhao, Chengyu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. A dual-way enhanced framework from text matching point of view for multimodal entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19008–19016, 2024.
39. Chengmei Yang, Bowei He, Yimeng Wu, Chao Xing, Lianghua He, and Chen Ma. Mmel: a joint learning framework for multi-mention entity linking. In *Uncertainty in Artificial Intelligence*, pages 2411–2421. PMLR, 2023.