VCRMNER: Visual Cue Refinement in Multimodal NER using CLIP Prompts

Yu Bai[1,2], Lianji Wang[1], Xiang Liu[1], Haifeng Chi[1], Guiping Zhang[1,2]

[1]School of Computer Science, Shenyang Aerospace University
[2]National and Local Joint Engineering Laboratory for Multilingual Collaborative Translation Technology
Correspondence: baiyu@sau.edu.cn

Abstract

With the continuous growth of multi-modal data on social media platforms, traditional Named Entity Recognition has rendered insufficient for handling contemporary data formats. Consequently, researchers proposed Multi-modal Named Entity Recognition (MNER). Existing studies focus on capturing the visual regions corresponding to entities to assist in entity recognition. However, these approaches still struggle to mitigate interference from visual regions that are irrelevant to the entities. To address this issue, we propose an innovative framework, Visual Cue Refinement in MNER(VCRMNER) using CLIP Prompts, to accurately capture visual cues (object-level visual regions) associated with entities. We leverage prompts to represent the semantic information of entity categories, which helps us assess visual cues and minimize interference from those irrelevant to the entities. Furthermore, we designed an interaction transformer that operates in two stages—first within each modality and then between modalities—to refine visual cues by learning from a frozen image encoder, thereby reducing differences between text and visual modalities. Comprehensive experiments were conducted on two public datasets, Twitter15 and Twitter17. The results and detailed analyses demonstrate that our method exhibits robust and competitive performance.

1 Introduction

Named Entity Recognition (NER) primarily identifies key entities (e.g., person, locations, organizations) within unstructured textual data sources (Li et al., 2020b). In the context of social media applications, NER technology is primarily used to analyze and track the dynamics of public opinion, major events, and other related information trends. As social networks evolve, the volume of multi-modal data on social media continues to grow, rendering traditional text-based NER methods insufficient for this new form of data. Consequently, researchers have developed Multi-modal Named Entity Recognition (MNER) (Lu et al., 2018)(Moon et al., 2018). MNER integrates image and text data to identify named entities, effectively resolving the ambiguities present in traditional NER tasks (Lu et al., 2018). It has now become an important research direction in the field of information extraction.

In MNER tasks, textual content and images often exhibit low relevance (Sun et al., 2021)(Hu et al., 2017), with entities usually concentrating on specific visual regions (visual cues). Other regions might interfere with the accurate identification of named entities (Xu et al., 2022)(Zhang et al., 2023a). Early studies (Lu et al., 2018)(Moon et al., 2018)(Wu et al., 2020)(Jia et al., 2022) have explored the inherent correlations between images and text using attention mechanisms. However, this approach does not address the low correlation between images and text, and it is challenging to assess the effectiveness of implicit alignments. Subsequently, research (Sun et al., 2021)(Xu et al., 2022) focused on reducing the influence of irrelevant images on entity recognition by evaluation mechanisms to assess the correlation between entire images and their corresponding textual sentences. For instance, they utilized contrastive learning methods to assess image-text similarity, or they employed pre-trained models for this assessment. Additionally, this approach diminishes the significance of object-level visual regions. Recently, studies (Chen et al., 2022)(Yu et al., 2023)(Zheng et al., 2020) have attempted to exploit object-level visual regions using visual tools such as Mask R-CNN (He et al., 2017). The object-level visual information typically corresponds directly to visual objects with less noise, these visual regions can better assist in entity recognition. However, these visual tools are typically trained solely on visual datasets, which hampers their ability to accurately capture the visual regions pertinent to the entities. Consequently,

Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025, pages 61–70
January 20th, 2025.

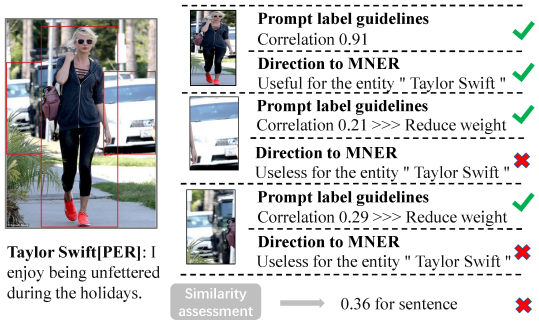

Figure 1: An example shows the problems of different MNER methods.

irrelevant visual regions can mislead the model's judgments, resulting in error propagation.

As shown in Fig. 1, the similarity score between the input image and the sentence, as evaluated by similarity assessment, is merely 0.36. Despite the person in the image being "Taylor Swift" as mentioned in the text, the relevance of the visual region where this person is located has decreased. In the method that utilizes Mask R-CNN to detect object-level visual regions within the image, three pertinent object regions were identified. However, the "car" in the image does not contribute to the identification of the entity "Taylor Swift". If the significance of this visual region is not diminished, it could potentially mislead the identification process of the entity.

Employing semantic information of entity category words to evaluate the relevance of visual regions helps minimize the interference caused by visual regions that are unrelated to the entities. CLIP (Radford et al., 2021) as a pre-trained multi-modal model, bridges the modalities of images and text. When used as an evaluator, CLIP is capable of assessing the relevance between images and textual content. However, the training corpus for CLIP lacks category words for entity classifications, including person, location, organization, and miscellaneous. Consequently, it is crucial to guide visual region assessments with category words.

In this paper, we propose a new framework Visual Cue Refinement for MNER (VCRMNER) that uses prompts instead of category words to guide the evaluation of visual regions. Specifically, we characterize the types of entities using a series of prompts and compute the center of these prompts' vector representations to represent the entity category words. The similarity between these category word representations and the representations of visual cues is assessed with CLIP to determine the relevance of these visual regions. Furthermore, we developed a two-stage modal fusion interactive transformer. First, attention is calculated within each modality separately. Then, the model fuses the modalities together. This balances the differences between text and visuals, avoiding excessive interference from visual representations. By refining visual cues from the frozen image encoder, our model reduces modal discrepancies effectively. Comprehensive experiments were conducted on two public datasets Twitter15 (Zhang et al., 2018) and Twitter17 (Lu et al., 2018). The results and detailed analyses confirm that our method provides robust and competitive performance. Our main contributions are summarized as follows:

- We proposed an innovative architecture VCRMNER that employs a transformer block that combines cross-attention and self-attention, interacting with a frozen visual encoder. This interaction reduces the semantic gap between modalities, thereby more effectively integrating information from different modalities to achieve MNER.
- We designed a prompt-guided visual cue evaluation module that supplements additional semantic information by using prompts to replace entity category words, thereby effectively reducing interference from visual noise unrelated to the entities.
- We conducted extensive experimental verification on two benchmarks, and the experimental results fully demonstrated that our method achieved sota.

2 Related Work

2.1 Prompt learning

The concept of prompt learning involves designing appropriate "prompts" to elicit the desired outputs from the model. The core of this approach lies the idea of not training the model directly for specific tasks, but rather constructing a form of input that enables the model to infer the correct answers based on existing knowledge (Liu et al., 2023). Recently, some studies (Huang et al., 2022) developed a prompt learning method for NER that uses category-specific words to optimize contrastive learning of label representations. This approach, however, is generally limited to the textual modality. In multimodal approaches, (Wang

et al., 2022) proposed a method that leverages the association between prompts and visual images to filter prompts containing entity semantics to assist in entity recognition. This method significantly mitigates the differences between the visual and textual modalities. In contrast, we propose using a specifically designed prompt-driven vision-language model as an evaluator, starting from raw data, to assess whether object-level visual regions (visual cues) are related to entity categories. This approach minimizes interference from irrelevant object-level visual regions and enhances entity recognition through precise visual cues.

2.2 Pretrain vision language model

With the continuous advancement of pre-trained models, significant progress has been made in the fields of computer vision and natural language processing. In this context, Unicoder (Li et al., 2020a) attempted to use a universal encoder to integrate visual and linguistic representations, drawing on the paradigm of cross-lingual pre-training models, inputting both visual and textual data into multi-layer Transformers for multi-task cross-modal pre-training, aimed at image-text retrieval tasks. CLIP (Radford et al., 2021), which uses large-scale image-text pairs and contrastive learning to predict the match between captions and images, thereby understanding the relevance between two different modalities. The CLIP model demonstrates strong generalization ability across various visual tasks without the need for specific task training.

These multimodal visual-language pre-trained models break down modality barriers, narrow the gap between modalities, and exhibit great potential in numerous downstream tasks. For the MNER task, images in image-text pairs often contain visual objects unrelated to entity types, and the existing multimodal visual-language models, with their capability to evaluate image-text associations, provide a technical foundation for the assessment of visual objects.

2.3 MNER

With the increasing amount of multi-modal data on social media platforms, MNER has attracted the attention of many researchers. Based on different image processing methods, we categorize MNER research into two main classes:

(1) Treating the entire image and merging through the interaction between image representations and text representations in vector space. For example, CNN-LSTM (Lu et al., 2018) introduces a modality attention module that diminishes irrelevant modality information while amplifying the primary modality, used for multi-modal representation. CoA (Zhang et al., 2018) introduces an adaptive co-attention architecture to integrate visual and textual information for MNER. UMT (Yu et al., 2020) has designed a unified multi-modal transformer framework that utilizes an entity span detection task to learn rich multi-modal representations. MAF (Xu et al., 2022) proposes a matching and alignment framework to mitigate the effects of mismatched text-image pairs and enhance the consistency of multi-modal representations. DebiasCL (Zhang et al., 2023a) employs implicit alignment between visual objects and textual entities, using debiasing-based contrastive learning to optimize the shared semantic space between text and images. These methods attempt to leverage entire image to enhance textual representations; however, they overlook the preference for object-level visual regions in the MNER task.

(2) Explicitly extracting object-level visual regions (visual cues) and facilitating interaction between representations of visual regions and textual representations. OCSGA (Wu et al., 2020) utilizes a dense co-attention mechanism to establish both intra-connections and inter-connections between textual entities and visual objects. UMGF (Zhang et al., 2021a) proposes a graph fusion method to learn various semantic relationships between words and multiple visual objects. BGAMNER (Chen et al., 2023) explores the matching relationships between visual regions and words through bidirectional image-text generation. The HamLearning (Liu et al., 2024) enhances text word representations by dynamically aligning image and text sequences and modeling the relationships among the mined visual regions, thereby achieving multi-level cross-modal learning. Although visual regions more accurately point to named entities, the visual regions in images are not always relevant to the entities.

Unlike the above methods, we propose an approach that uses prompts instead of entity category words to evaluate visual cues and dynamically achieves word alignment with visual cues and inter-modal fusion through an interactive transformer.

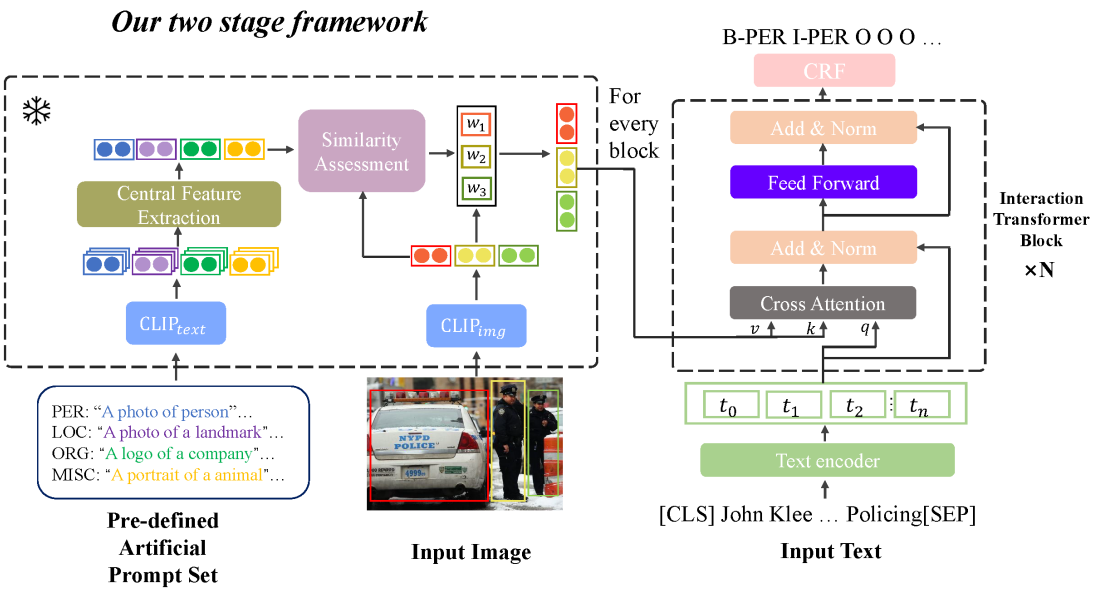

Figure 2: The overall architecture of the approach we proposed.

3 Our Method

Overview our method. Fig. 2 shows our model architecture, which is divided into two main stages. In the first stage, we design the corresponding ten text prompts for each entity category word and map each text prompt to the multi-modal vector space through the clip text encoder. Further, we find the center of the vector representation of the prompts corresponding to each entity category by averaging, and we use this center as the text vector representation of the entity category words. Subsequently, we map the object-level visual regions, obtained by Mask R-CNN, to the shared vector space of image and text via the clip. We then compare the visual representation with the prompts center to evaluate the correlation. In the second stage, we designed an interactive transformer aimed at refining the captured visual cues, enabling the model to more effectively learn the knowledge embedded in the visual modality. This module achieves sustained intra-modal and inter-modal interactions through self-attention and cross-attention mechanisms, thereby narrowing the semantic gap between the textual and visual modalities. Finally, we input the final vector representation obtained from the interaction transformer into the Conditional Random Field (CRF) (Wallach et al., 2004) to complete the sequence annotation decoding process. The following section describes the details of our method.

Formula definition. Given a sentence $S = (w_1, w_2, w_3, ..., w_n)$, where n represents the total number of words in the sentence. I is the image corresponding to the sentence S. The goal of the MNER task is to combine the image I to label each word in the sentence S to obtain the sequence $\hat{Y}$. $\hat{y}_i \in \hat{Y}$, $\hat{y}_i$ type is the label of the model predict, $Y = (y_1, y_2, y_3, \ldots, y_n)$, y_i represents the type of the real label. According to the BIO notation, the MNER task pre-defines four types of entities $PER, LOC, ORG, MISC$.

Input Embedding: In this paper, we choose RoBERTa (Liu et al., 2019) as the text encoder. Before feeding a sentence S into RoBERTa, the RoBERTa tokenizer segments it into a sequence of word embeddings. Special tokens "[CLS]" and "[SEP]" are inserted at the beginning and end of the word embedding sequence, respectively. This process generates a token sequence $T = \{T_i\}_{i=1}^{N_t} \in \mathbb{R}^{N_t \times d}$, where N_t represents the number of tokens in the sentence, and d represents the dimension of the embeddings. The token sequence is then input into RoBERTa to obtain the vector representation $H_t \in \mathbb{R}^{N_t \times d}$, where h_0 is the vector representation of the entire sentence, and the others are the vector representations of the words in the sentence.

As a Transformer-based image recognition model, Vision Transformer (ViT) (Dosovitskiy et al., 2020) efficiently processes global features of images through the self-attention mechanism.

Therefore, we choose clip-vit as the model's visual encoder. Given an image I, ViT segments the input image into multiple fixed-size image patches (e.g., 16x16 pixels) and linearly projects these image patches into a series of one-dimensional embedding vectors, represented as $V = \{V_i\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times d}$, where N_v represents the number of patches in the image, and d represents the embedding dimension.

3.1 Stage one. Prompt-guided visual cues extraction and evaluation

Prompts for design: When evaluating the relevance of object-level visual regions, directly using label words from named entity recognition (such as person, location, and organization) may lead to low recognition accuracy because these keywords do not directly correspond to the content in the CLIP model's training set. To overcome this limitation, we introduce a prompt-based visual cue method that utilizes a set of carefully designed prompts as substitutes for entity category words, thereby harnessing CLIP's capability to assess visual regions. We designed ten distinct prompts for each entity category using the large language model GPT-4, which generated and screened prompts based on the MNER task description and the definitions of entity category words. These prompts were then manually filtered to select those that accurately describe the image content and align with the entity type definitions. All prompts are available in the supplementary material.

Visual cues assessment: The predefined prompts are used in place of entity keywords and are encoded through a CLIP text encoder, which converts each prompt into a vector representation. Given $prompt_i^j$ has the following formula to map it to a multi-modal vector representation space:

$$p_i^j = clip_{text}(prompt_i^j) \in \mathbb{R}^d \tag{1}$$

where $p_i^j \in \mathbb{R}^d$ represents the vector embedding of the i-th prompt in the j-th category, with d denoting the dimensionality of the embedding space, and $clip_{\text{text}}$ is the CLIP model's text encoder, responsible for mapping the input prompt into the shared multi-modal vector space.

After obtaining the vector representation of all prompts for each category, we obtain the entity category textual representation by calculating the average of these vectors:

$$\mathbf{c}^j = \frac{1}{n}\sum_{i=1}^{n} p_i^j \tag{2}$$

where $\mathbf{c}^j \in \mathbb{R}^d$ represents the centroid vector of the j-th entity category in the multi-modal embedding space, obtained by averaging the n prompt vectors p_i^j, with n denoting the total number of prompts in the category and d being the dimensionality of the embedding space.

It is noteworthy that, since the entity category words are fixed, the prompts we design will also be fixed. Therefore, the process of computing the center of the prompt is performed only once. After obtaining the vector representation of the prompts center, it is passed as a fixed parameter into the model. As the text prompts undergo only a single pass through the text encoder and are fixed, the model we propose does not lead to excessive computational growth.

Given an image, we follow the approach of (Zhang et al., 2021b) by utilizing the Visual toolkit to extract the top m most salient local visual objects. These visual cues $O = \{O_1, o_2, o_3, ..., o_m\}$ are then resized to 224×224 pixels and mapped into the multimodal representation space using the CLIP visual encoder. The process is represented as follows:

$$v_i' = clip_{img}(o_i) \tag{3}$$

where o_i represents the visual objects. v_i' denotes the vector representation of the i-th object-level visual region in the multimodal vector space.

Subsequently, the text representation of entity category words c^j and the visual representation v_i' are normalized to eliminate the influence of the length of the vector on the similarity calculation so that the similarity is mainly affected by the direction of the vector and not by its length.

$$c^j = \frac{c^j}{D_c}, v_i = \frac{v_i}{D_v} \tag{4}$$

where D_c and D_v are the dimensions of c^j and v_i, and then we use softmax to increase the disparity in similarities.

$$w\_sim_i^j = softmax(logit * v_i * c^{jT}) \tag{5}$$

where $w\_sim_i^j$ represents the relevance between the i-th visual region and the j-th class entity, $logit$ is a parameter in clip which utilized to enhance the discriminative power between categories.

After obtaining the similarities between visual regions and various categories, we select the similarity $w\_sim_i^j$ with the maximum relevance to the visual region as the weight to update the visual representation. At this point, visual regions that are

irrelevant or have low relevance are assigned lower weights.

$$cue_v = cat(v_i * max(w\_sim_i^j), 2) \quad (6)$$

where cue_v is the visual cues, and cat is the concatenate method that concatenates each visual representation along the second dimension. cue_v is the visual representation fed into the interaction transformer.

3.2 Stage two. Interaction transformer for visual cues refinement

To refine and effectively learn visual cues and achieve interactions and fusion between modalities, we have designed an interaction transformer architecture that interacts with the visual representations obtained from a frozen visual encoder. Within this framework, textual interacts with context through self-attention mechanisms to capture intra-modal semantic associations. The updated textual representations are then fed into alternating cross-attention as queries, thereby facilitating the extraction of associations between text and visual modalities.

In the intra-modal interaction process, self-attention is first used to learn the associative information within the modality. Given the input sentence S, we utilize text encoder to obtain the textual representation:

$$H_t^0 = encoder_t(S) \quad (7)$$

where $H_t^0 \in \mathbb{R}^{N_t \times d}$ is the textual representation. Subsequently, for intra-modal interactions, we employ the classical multi-head self-attention mechanism to update the textual representations. During this computation process, the updated textual context representations are calculated as follows:

$$Q = H_t^{l-1}W^q, K = H_t^{l-1}W^k, V = H_t^{l-1}W^v \quad (8)$$

$$I_t^{'} = MA(Q, K, V) \quad (9)$$

where l denotes the number of the layers, $I_t^{'}$ represents the representation after intra-modal interactions, $MA(Q, K, V)$ is the multi-head attention mechanism, and Q, K, V are the query matrix, key matrix, and value matrix respectively. $I_t^{'}$ is directly input into the feedforward network:

$$I_t = LN(I_t^{'} + FFN(I_t^{'})) \quad (10)$$

After completing intra-modal interactions, we employ a cross-attention mechanism to facilitate inter-modal interactions and refine visual cues, thereby reducing the differences between modalities. With I_t serving as the Query in cross-attention calculations with the visual representations. Given the visual representation cue_v. The computation process is as follows:

$$Q = I_t^{l-1}W^q, K = cue_vW^k, V = cue_vW^v \quad (11)$$

$$I_m^{'} = CA(Q, K, V) \quad (12)$$

$$I_m = LN(I_m^{'} + FFN(I_m^{'})) \quad (13)$$

where I_m represents the representation after intra-modal interactions, $LN()$ is the layer norm, $CA()$ is the cross attention, Q, K, V are the query matrix, key matrix, and value matrix respectively.

3.3 Label prediction and Model training

A decoder is required to decode the final representations. Extensive research has demonstrated the superior performance of CRF in sequence labeling tasks. CRF is capable of extracting hierarchical information from the semantic space for sequence labeling and has achieved commendable results in numerous sequence labeling tasks. Consequently, we employ a CRF for the purpose of decoding.

$$P(y \mid G) = \frac{\prod_{i=1}^{n} E_i(y_{i-1}, y_i, G)}{\sum_{y' \in Y} \prod_{i=1}^{n} E_i(y'_{i-1}, y'_i, G)} \quad (14)$$

where G is the final vector representation output by the interactive transformer. We choose the maximum likelihood function to calculate the loss and train our model:

$$\mathcal{L} = -\sum_{j=1}^{M} \left(\log P(y^j \mid G^j)\right) \quad (15)$$

4 Main Result

4.1 Experimental Setup

Dataset. Our experimental tests are consistent with previous studies using two benchmarks: Twitter15 (Lu et al., 2018) and Twitter17 (Zhang et al., 2018). Table 2 shows the basic statistics of the two benchmarks.

Table 2: The basic statistics of twitter15 and twitter17.

Entity Type	twitter15			twitter17		
	Train	Dev	Test	Train	Dev	Test
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	927	247	839	1674	375	395
Miscellaneous	931	220	720	701	150	157
Total	6166	1541	5072	6049	1324	1351
Num of sentence	4000	1000	3257	3273	723	723

Table 1: The proposed method was compared with several baselines on the Twitter15 and Twitter17 benchmark datasets.

Modality	Methods	twitter15			twitter17		
		P	R	F1	P	R	F1
TEXT	CNN-BLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	HBiLSTM-CRF	70.30	68.05	69.17	82.69	78.16	80.37
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
TEXT+IMAGE	AdapCoAtt(Zhang et al., 2018)	69.87	74.59	72.15	85.13	83.20	84.10
	OCSGA(Wu et al., 2020)	74.71	71.21	72.92	-	-	-
	RpBERT(Sun et al., 2021)	71.15	74.30	72.69	-	-	-
	UMT(Yu et al., 2020)	71.67	75.23	73.41	85.28	85.34	85.31
	UMGF(Zhang et al., 2021a)	74.49	75.21	74.85	86.54	84.50	85.51
	MEGA(Zheng et al., 2021)	70.35	74.58	72.35	85.84	87.93	86.87
	HVPNeT(Chen et al., 2022)	73.87	76.82	75.32	85.84	87.93	86.87
	MAF(Xu et al., 2022)	71.86	75.10	73.42	86.13	86.38	86.25
	DebiasCL(Zhang et al., 2023a)	74.45	76.13	75.28	87.59	86.11	86.84
	TGF(Zhang et al., 2023b)	73.88	75.98	74.91	**88.42**	86.96	87.70
	BGA-MNER(Chen et al., 2023)	**78.6**	74.16	76.31	87.71	87.71	87.71
	HamLearning(Liu et al., 2024)	77.25	75.75	76.49	86.99	87.28	87.13
	VCRMNER(Ours)	75.48	**78.23**	**76.83**	87.76	**89.79**	**88.76**

Implementation details. Our experiments were conducted under one Nvidia Tesla T4, using the PyTorch 1.8.0 framework to build the model. We use roberta (Liu et al., 2019) base as the text encoder and clip base as the visual encoder and evaluator. The visual encoder was frozen. The learning rates for the Interaction Transformer and the text encoder were set at 4e-5, while the learning rate for the CRF was established at 1e-4. We employed a linear warm-up strategy, with the warm-up rate set at 1e-2. Within the model, the heads of multi-head attention were set to 8. Interaction transformer blocks were configured 3. The maximum sequence length for the text was determined to be 70, ensuring coverage of all words within the sentences. The model was trained over 40 epochs with a batch size of 18.

4.2 Main Experimental Results and Analysis

To validate the effectiveness of the proposed method VCRMNER, we select a total of 14 baseline methods for comparison, including both pure text-based approaches and multimodal methods. In our experiments conducted on the Twitter15 and Twitter17 datasets, we employed precision (P), recall rate (R), and F1 score (F1) as evaluation metrics. We compared our method with several competitive MNER methods. The results in Table 1 demonstrate that our method has outperformed the current sota methods.

Firstly, under the single-text modality, the method of fine-tuning a pre-trained language model demonstrates significant advantages over the approach using a non-pretrained BiLSTM model. This indicates that the rich prior knowledge embedded in pre-trained models plays a crucial role in the task of NER, thereby enhancing the recognition performance.

Secondly, by comparing methods between multimodal and single-text modalities, we found that approaches utilizing either entire image or visual regions consistently outperformed those relying solely on the single-text modality. These results adequately demonstrate the importance of visual information in MNER tasks. Furthermore, methods that utilize visual regions, such as HVP (Chen et al., 2022), BGA-MNER (Chen et al., 2023), and TGF (Zhang et al., 2023b), have shown clear advantages over those using entire images, like UMT (Yu et al., 2020) and UMGF (Zhang et al., 2021a).

Lastly, methods such as MAF (Xu et al., 2022) and DebiasCL (Zhang et al., 2023a), which assess the significance of visual images, have proven to be crucial in enhancing the effectiveness of MNER tasks, as evidenced by experimental results. This underscores the indispensability of evaluating visual regions in MNER tasks. Compared to methods that assess relevance using sentences and entire image, such as MAF (Xu et al., 2022), DebiasCL

Table 3: Results of ablation study for the MNER task.

Method	twitter15			twitter17		
	P	R	F1	P	R	F1
w/o Inter-former	**75.68**	76.46	76.06	87.44	87.64	87.54
w/o ev	74.64	76.21	75.42	86.26	88.75	87.49
w/o prompt	74.17	77.96	76.02	87.2	89.27	88.22
VCRMNER(Ours)	75.48	**78.23**	**76.83**	**87.76**	**89.79**	**88.76**

Table 4: Results of cross-domain performance on different methods.

cross-domain	17→ 15			15→ 17		
	P	R	F1	P	R	F1
HamLearning	69.17	66.84	67.98	71.03	59.4	64.7
BGA-MNER	**72.17**	67.98	68.71	70.81	59.6	64.91
VCRMNER(Ours)	71.87	**68.31**	**69.08**	**72.45**	**60.25**	**65.11**

(Zhang et al., 2023a), and HamLearning (Liu et al., 2024), our approach has demonstrated significant advantages. This further confirms the critical importance of precise evaluation of visual regions in MNER tasks.

4.3 Further experiments and analysis

Ablation Study. To explore the effectiveness of each component of our proposed method, we conducted comprehensive ablation studies. The results of these studies are shown in Table 3, where "Interformer" refers to the interaction transformer module, "ev" denotes our initial evaluation of visual regions, and "prompt" indicates our method of assessing visual regions using entity category words instead of prompts.

The experimental results demonstrate that every component of our proposed method is effective; removing any part leads to a decrease in model performance. The most significant decline in performance occurs when the evaluation module is ablated, highlighting the critical importance of visual cue assessment in MNER. Eliminating the method of using entity category words for assessment in favor of prompts significantly reduces performance, indicating that prompts align more closely with visual representations than do entity category words. By ablating the Interaction Transformer module and directly concatenating text and visual representations, the lack of effective inter-modal interaction leads to a substantial disparity between visual and textual modalities, resulting in decreased performance.

Cross-domain generalizability analysis. We swapped the test sets of Twitter15 and Twitter17. For instance, we trained on Twitter17 and tested on Twitter15. From Table 4, we observe that our model maintains competitive performance across various test sets. Compared to previous state-of-the-art methods, our model demonstrates certain advantages, indicating the strong generalization capability of the proposed approach.

5 Conclusion

In this paper, we proposed a new framework to implement the MNER task in two stages. It guides Mask R-CNN to mine visual objects through prompts and obtains visual objects closely related to the entity category words. Through the interactive transformer, we refined the visual cues and narrowed the semantic gap between modalities. We have constructed a variety of experiments to prove that our method is effective and achieves SOTA effects.

6 Limitation

While we have placed the calculation of prompt centers outside the training process to avoid excessive increases in computational complexity during model training and inference, the necessity of evaluating visual cues requires us to employ the CLIP text encoder. Consequently, the final model includes the CLIP text encoder, inevitably leading to an increase in the overall number of model parameters.

References

Feng Chen, Jiajia Liu, Kaixiang Ji, Wang Ren, Jian Wang, and Jingdong Chen. 2023. Learning implicit

entity-object relations by bidirectional generative alignment for multimodal ner. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4555–4563.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. 2017. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International conference on computational linguistics*, pages 2515–2527.

Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang, Lejian Liao, Yang Song, Meng Chen, and Xiaodong He. 2022. Query prior matters: a mrc framework for multimodal named entity recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3549–3558.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020b. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Peipei Liu, Hong Li, Yimo Ren, Jie Liu, Shuaizong Si, Hongsong Zhu, and Limin Sun. 2024. Hierarchical aligned multimodal learning for ner on tweet posts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18680–18688.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.

Hanna M Wallach et al. 2004. Conditional random fields: An introduction. *University of Pennsylvania CIS Technical Report MS-CIS-04-21*, 24:33–42.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022. Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *International Conference on Database Systems for Advanced Applications*, pages 297–305. Springer.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: a general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1215–1223.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021b. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023a. Reducing the bias of visual objects in multimodal named entity recognition. In *Proceedings of the Sixteenth ACM international conference on web search and data mining*, pages 958–966.

Zhengxuan Zhang, Weixing Mai, Haoliang Xiong, Chuhan Wu, and Yun Xue. 2023b. A token-wise graph-based framework for multimodal named entity recognition. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2153–2158. IEEE.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532.

A Appendix

Prompts Used in the Experiment:

The prompts employed in our experiment were categorized into four main groups: *Person (PER)*, *Location (LOC)*, *Organization (ORG)*, and *Miscellaneous*. Each category was designed to capture a specific aspect of visual content, facilitating a comprehensive analysis across diverse image types.

Person (PER)

In the *Person* category, we included a range of human subjects. This involved images such as a photo of a person, an image of a woman, and a photo of a child. Additional variations included a picture capturing someone, an image of a person with glasses, a portrait of an elderly man, a snapshot of a teenager, a group photo of a family, a candid shot of a person laughing, and a studio portrait of a young adult. These selections aimed at representing different age groups, genders, and social contexts.

Location (LOC)

The *Location* category encompassed various geographical and architectural elements. It featured a photo of a famous landmark, a scenic view of a well-known city, the landscape of a famous natural wonder, a street view in a recognizable city, the architecture of a well-known building, a panoramic view of a historic site, a night shot of a city skyline, a sunrise behind famous city landmarks, a detailed architectural close-up of a historical building, and a picturesque view of a village. This variety ensured that both urban and natural settings were adequately represented.

Organization (ORG)

For the *Organization* category, we focused on institutional and corporate imagery. This included the exterior of a famous institution, a logo of a well-known company, the entrance of a renowned university, a branded product from a famous manufacturer, an official sign of a governmental organization, the front view of an international airport, the headquarters of a global tech company, a franchise store of a popular brand, the emblem of a prestigious college, and a product lineup of a leading electronics brand. These images were chosen to reflect the diversity of organizational structures and their public representations.

Miscellaneous

Lastly, the *Miscellaneous* category covered a wide array of objects and scenes not fitting into the previous categories. This included a close-up photo of a consumer electronic device, a portrait of an animal, an image depicting a traditional cultural festival, a detailed image of a plant, a macro shot of a unique flower, a still life photo of a classical instrument, an artistic depiction of a folk dance, a photo of intricate jewelry, a high definition image of an exotic bird, and a festive scene from a national holiday. The aim here was to introduce a broader spectrum of visual interests and cultural elements.