MGIPS-CL: Multimodal NER with Multi-Granularity Interweave and Implicit Positive Sample Contrastive Learning

ARTICLE INFO

Keywords:
Information Extraction
Multimodal Named Entity Recognition
Multi-Granularity Interweave
Mutual Attention
Implicit Positive Sample Contrastive Learning

ABSTRACT

In the task of Multimodal Named Entity Recognition (MNER), visual information has been shown to significantly enhance recognition performance. However, existing methods predominantly rely on co-attention mechanisms to model the relationship between text and images, which still face two major limitations: 1) Insufficient exploitation of potential semantic relevance across text-image pairs, leading to limited recognition capability; 2) Single-granularity fusion strategies that either lose important information or introduce excessive noise, thereby hindering deep cross-modal integration. To address these problems, we propose MGIPS-CL, a novel MNER framework based on Multi-Granularity Interweave and Implicit Positive Sample Contrastive Learning. We utilize Implicit Positive Sample Contrastive Learning (IPS-CL) to achieve implicit alignment of two modalities. It shuffles the relationships of the original text-image pairs and selects the most relevant image for the text to dynamically constructing an Implicit Positive Sample Pair (IPSP) based on the potential relevance of the image to the text. And we propose a Multi-Granularity Interweave Module (MGI) which performs hierarchical cross-modal fusion at different interweaving levels at a multi-granularity way. It can effectively filter noise while reducing the loss of crucial information. We conduct experiments on the real-world multimodal social media datasets (Twitter-2015 and Twitter-2017). The results demonstrate that our method outperforms the SOTA model, achieving F1-scores of **76.86**% and **88.52**%, respectively.

1. Introduction

The Named Entity Recognition (NER) has attracted increasing attention as the field of Natural Language Process (NLP) evolves. NER [18, 42] is aim to identify specific types of entities, such as Person (PER), Location (LOC), and Organization (ORG). In NLP [20], NER has been extensively utilized in domains such as news, e-commerce, and social media.

In certain cases, textual information may not offer enough contextual information to accurately identify the entity type. For instance, in the sentence "Allen was drenched in the rain", it is challenging for a text-based NER to determine whether "Allen" refers to a human or a dog. To address this issue, we integrate visual information, as shown in Fig.1(a), to enhance the accuracy of Multimodal Named Entity Recognition. Consequently, "Allen" can be correctly identified as a dog.

Some methods [50, 45, 23, 4] utilize RCNN for object detection in images to establish explicit alignment between the objects and entities in the text. For instance, the image in Fig.1(b) precisely represents the text content and distinctly displays three PER type objects. It is easy to determine that they are all belong to the PER type. However, in Fig.2(b), the information that "Obama and Trump are person" is missing. This means that the pair of image and text does not match. It becomes challenging to judge whether Obama and Trump belong to the PER type without referring to Fig.2(a) and Fig.2(c). According to [30], in over 33.8% of tweets, the images are completely unrelated to the text, indirectly suggesting that in more tweets, the images do not entirely correspond to the text. In other words, solely relying on a single image cannot accurately determine the entity type in the associated text. Therefore, it is necessary to use other related images as a supplement, and this is the first challenge we face.

To address the above challenge, we design a novel Implicit Positive Sample Contrastive Learning (IPS-CL), which mines Implicit Positive Sample Pairs (IPSP) and fully utilizes the semantic information between IPSP to achieve implicit alignment. By shuffling the original pairing relationships between text-image pairs, we calculate the potential relevance coefficient between text and each image, then we select K images with the highest relevance for each text to form K groups of text-image pairs as the positive sample pairs in the contrastive learning, which we call IPSP.

Another methods [46, 43, 35] utilize the gate mechanism to directly integrate image and text features. However, they can't effectively filter the noise brought by images. And there are also some methods utilize the bottleneck fusion method [32, 38] to fuse in a single-granularity way. While it effectively filters noise, its narrow focus on fine-grained

ORCID(s):

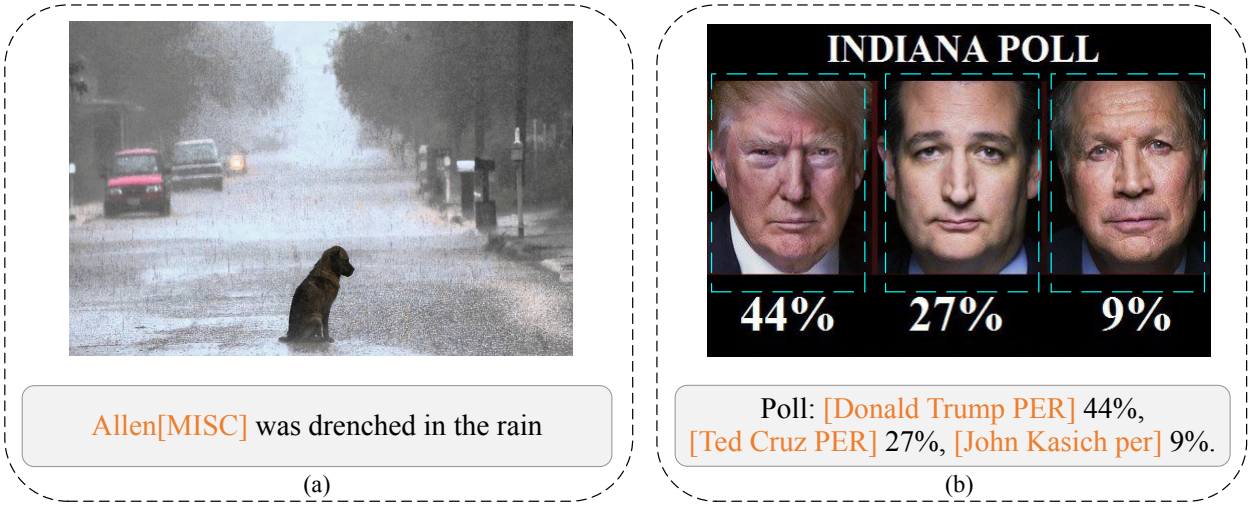


Figure 1: Fig (a) is an example for MNER. Fig (b) is an example for MNER which contains three objects.

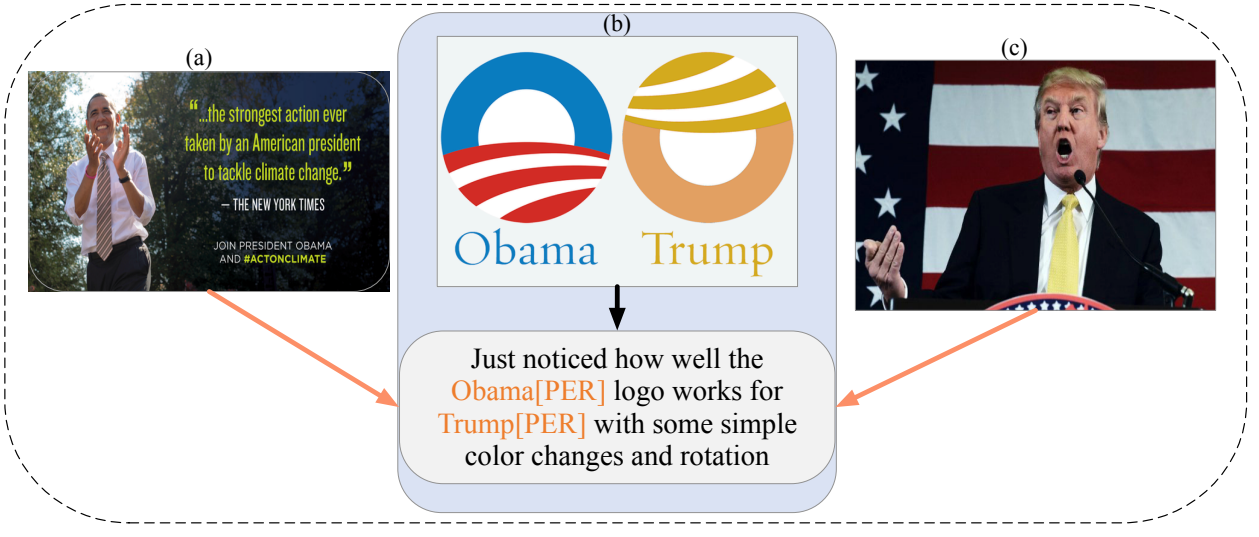


Figure 2: An example for MNER without objects.

fusion can result in the loss of crucial coarse-grained information, leading to poor model generalization. So how to effectively filter noise while reducing the loss of crucial information is the second challenge we face.

To address the second challenge, we design the Multi-Granularity Interweave Module (MGI). First, We use Interweave Encoder to interweave textual and visual features with different degree of interweaving. The Multi-Granularity Extractors (MGE) guided by image or text share the same encoder, respectively. And because textual and visual features with different degree of interweaving have different aspects of information, MGE is able to encode multi-granularity features with richer information. Then we use pre-post adaptive gate to combine the post-fusion textual features with the pre-fusion textual features rather than directly combining the image and text for the classification task. The pre-post adaptive gate is similar to the Gated Multimodal Unit proposed by [1], which has been proven to be effective in fusing information. We modify it slightly to suit our task. In this way, MGI effectively mine rich multi-granularity information through Interweave Encoder while reducing noise propagation through indirect use of visual features.

To this end, we effectively combine IPS-CL and MGI to propose an MNER method with Multi-Granularity Interweave and Implicit Positive Sample Contrastive Learning (MGIPS-CL), which can effectively solve the above problems.

Our main contributions are listed below:

- We design a Multi-Granularity Interweave Module for MNER, which effectively interweave visual and textual features at a multi-granularity way, enabling deep fusion of two modalities.
- We propose an Implicit Positive Sample Contrastive Learning that can mine Implicit Positive Sample Pairs based on the potential relevance between text and images and effectively utilize the relevant information in the Implicit Positive Sample Pairs.
- We conduct comprehensive experiments and sensitivity analysis on twitter-2015 and twitter-2017 datasets, and the results show that our model outperforms the SOTA model, on the Twitter-2015 and Twitter-2017 datasets, achieving F1-scores of **76.86**% and **88.52**%, respectively.

2. Related Work

2.1. Multimodal Named Entity Recognition

In the era of booming social media, MNER has been received more and more attention. [25] proposed LSTM-CNN hybrid multimodal model. [39] innovatively combined the Mask RCNN [46] for visual object recognition with co-attention. Thereafter, [43] designed a multimodal interaction module to capture semantic relationships between words and images. [31, 48, 33, 34] Utilized innovative methods such as heterogeneous graphs, scene graphs, image-text alignment, and prompt learning to enhance the recognition ability of named entities. In addition, [46] proposed an adaptive co-attention and gated fusion mechanism to fuse features of different modalities directly. [12] uses entity types and visual regions to enhance the representation of both texts and images. [36] designs the Correlation-Aware Alignment layer, combined with contrastive learning, to achieve implicit alignment between modalities.

However, not all images are closely associated with textual information in social media data. Fusing image and text information at a single-granularity may introduce more noise. To this end, we use Multi-Granularity Interweave Module to filter noise from images while realizing cross-modal information fusion.

2.2. Contrastive Learning

Contrastive learning has achieved remarkable success and attracted significant attention in the fields of Computer Vision (CV) and Natural Language Process (NLP). It brings significant improvements to a variety of downstream tasks. In [10, 11], positive sample pairs can be from different pairs, different samples that have the same label, or different forms of the same sample based on data augmentation [5]. In the field of CV, [5, 15] have utilized image augmentation to generate positive samples and achieved positive results. While in the field of NLP, [8, 40] searched for appropriate text augmentation methods to provide positive samples by text augmentation. The introduction of InfoNCE loss [29] plays a significant role in downstream tasks [29, 5, 9, 14]. With the increasing availability of multimodal data, some researches [19, 27, 17] have begun to extend contrastive learning to the multimodal domain. Most of them utilize standard contrastive learning which takes the original image-text pairs in the same batch as the positive sample pairs.

However, standard cross-modal contrastive learning methods ignore the problem of mismatch between text-image pairs and the potential relevance between different text-image pairs, which will affect the performance of the model to some extent. To address this issue and fully leverage the performance of the model, we introduce Implicit Positive Sample Contrastive Learning which mines different text-image pairs with high potential relevance as IPSP for contrastive learning.

2.3. Multimodal Fusion

The image pyramid approach [21] has been widely utilized to enrich visual features by fusing features of different layers. Recently, some researchers have proposed to use bottleneck fusion [38, 32] for inter-modal fusion in a fine-grained way and achieved positive results. [36] designs the Correlation-Aware Deep Fusion layer to achieve deep semantic fusion between modalities. Inspired by them, we propose Multi-Granularity Interweave Module (MGI). The textual and visual features with different interweaving degrees are input into the MGI to fuse at a multi-granularity way to mine textual features interweaved with visual features. Subsequently, we perform pre-post adaptive gate on the

textual features before and after fusion. In this way, we can not only mine the textual features fused with visual features but also effectively suppress the propagation of noise in visual features.

3. Method

In this section, we elaborate the details of each module in MGIPS-CL. Fig.3 shows the overall architecture of the models.

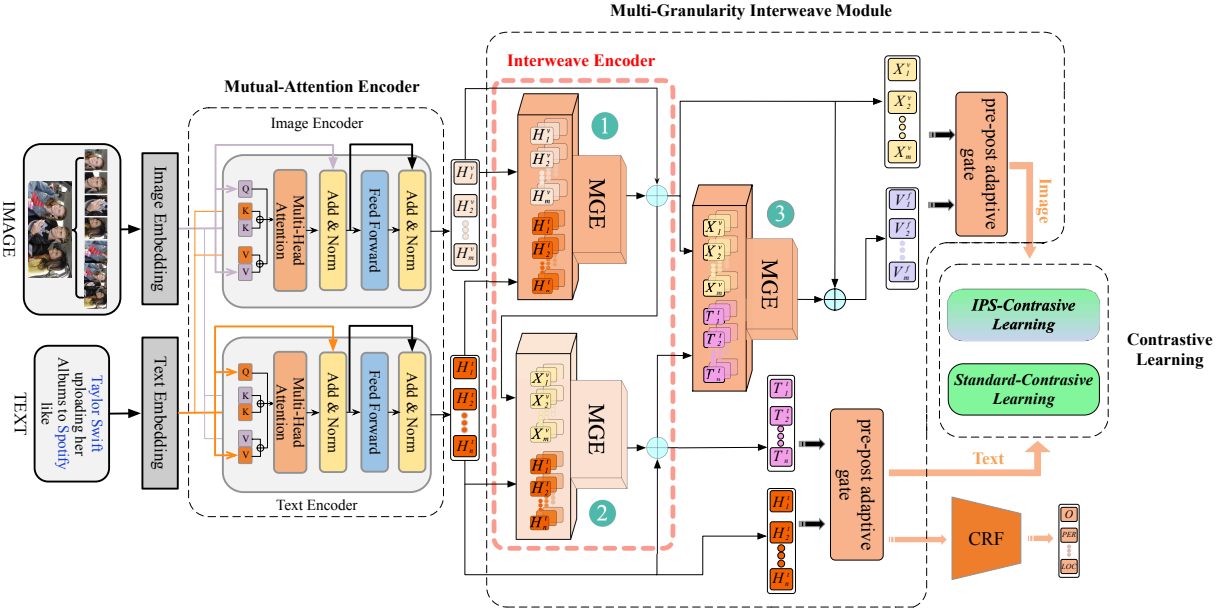


Figure 3: The Overall Architecture of Our MGIPS-CL. The MGE labeled 1 and 3 are image-guided MGE that share the same encoder, the MGE labeled 2 is text-guided MGE. Interweave Encoder consists of image-guided MGE and text-guided MGE.

3.1. Textual and Visual Embedding

3.1.1. Textual Embedding

We use the BERT Segmentation Tool to segment the input sentence to get $S = [s_1, s_2, ..., s_n]$, which denotes the tokens of the sentence after segment, n denotes the length of the tokens, s_i denotes the ith token, "[CLS]" and "[SEP]" are inserted tokens to represent the begin and the end of the sentence, respectively, and summed up with the position embedding and segment embedding, and then obtain $S^e = \{S_1^e, S_2^e, ..., S_n^e\}$.

$$S^e = S + S_{position} + S_{segment}. \tag{1}$$

3.1.2. Visual Embedding

We use the images obtained by visual grounding[1], each original image corresponds to three visual grounding images, and each image size is 128×128. We divide the images into patches of size 32×32 to obtain 3×4×4 patches. Similar to the "[CLS]" token in BERT, we concat a learnable embedding class embedding and sum it with the position embedding to get visual embedding $V^e = \{v_0^e, v_1^e, ..., v_m^e\}$, where m denotes the number of image patches, v_0^e denotes the class embedding, and v_i^e denotes the representation of the ith patch.

3.2. Mutual-Attention Encoder

We employ a transformer-based Multimodal approach to capture the correlations between visual objects and entities. Given the output $H_l^t \in R^{n \times d}$ of the lth layer in the text encoder. $Q_{l+1}, K_{l+1}, V_{l+1}$ are the key, value, query of

[1] Visual Grounding aims to find relevant objects or regions of a picture using natural language queries.

the text encoder at layer l+1 respectively, and we define it as:

$$\begin{aligned} Q_{l+1}^{t} &= H_{l}^{v} W_{l+1}^{q}, \\ K_{l+1}^{t} &= [K_{l}^{v}, H_{l}^{l} W_{l+1}^{k}], \\ V_{l+1}^{t} &= [V_{l}^{v}, H_{l}^{l} W_{l+1}^{v}], \end{aligned} \tag{2}$$

where $[\cdot, \cdot]$ denotes the concatenation, K_{l}^{v}, V_{l}^{v} denotes the key and value of the lth layer in the image encoder, and $W_{l+1}^{q}, W_{l+1}^{k}, W_{l+1}^{v}$ denote the projection parameters of the attention. In particular, when $l = 0$, $H_{0}^{t} = S^{e}$, so there is:

$$\begin{aligned} Q_{1}^{t} &= S^{e} W_{1}^{q}, \\ K_{1}^{t} &= [V^{e}, S^{e} W_{1}^{k}], \\ V_{1}^{t} &= [V^{e}, S^{e} W_{1}^{v}]. \end{aligned} \tag{3}$$

The output of the $l + 1$th layer is expressed as:

$$\begin{aligned} H_{l+1}^{t} &= Attn(Q_{l+1}^{t}, K_{l+1}^{t}, V_{l+1}^{t}), \\ H_{l+1}^{v} &= Attn(Q_{l+1}^{v}, K_{l+1}^{v}, V_{l+1}^{v}), \end{aligned} \tag{4}$$

where H_{l+1}^{t} denotes the output of the $l + 1$th layer in the text encoder and H_{l+1}^{v} denotes the output of the $l + 1$th layer in the image encoder.

3.3. Multi-Granularity Interweave Module

Our first innovation is Multi-Granularity Interweave Module (MGI). In this section we will introduce its three important parts: Multi-Granularity Excavator, Interweave Encoder and Pre-Post Adaptive Gate.

MGI can simultaneously interweave features of two modalities in both Row-grained and Column-grained ways and interweave features with different degrees of interweaving to mine richer information. By constraining the flow of visual features within the MGI framework, we reduce the propagation of noise caused by the image and coarse-grained. Furthermore, we integrate the textual features before and after fusion by pre-post adaptive gate, further ensuring that noise is effectively mitigated while preserving valuable information.

Note that the number of image-guided MGE is one more than the number of text-guided MGE. The visual features extracted in this way incorporate more comprehensive textual information, enabling us to accurately identify Implicit Positive Sample Pairs (IPSP) and better utilize Implicit Positive Sample Contrastive Learning (IPS-CL).

3.3.1. Multi-Granularity Excavator

In order to mine the semantic fusion features between two modalities, MGIPS-CL introduces MGE that takes into account both Row-grained and Column-grained. Specifically, we use the image-guided MGE as an example, as shown in Fig.4. Firstly, we calculate the intermediate matrix M_{im}:

$$M_{im} = (f_{1}(Linear_{1}(H^{t}))) \odot (f_{2}(Linear_{2}(H^{v}))), \tag{5}$$

where f_{1} and f_{2} are two encoders that convert H^{t} and H^{v} into a unified dimension, respectively. We then calculate the weights for the Row-grained and Column-grained aspects separately:

$$\begin{aligned} M_{C} &= softmax(M_{im} \otimes M_{3}, dim = -2), \\ M_{F} &= softmax(M_{im} \otimes M_{4}, dim = -1), \end{aligned} \tag{6}$$

where $M_{3} \in R^{d \times 1}$, $M_{4} \in R^{d \times 1}$, M_{C} denotes the weight in Row-grained, M_{F} denotes the weight in Column-grained, $\otimes$ denotes dot product. The weighted representation can be obtained by multiplying H^{v} with M_{F} and M_{C} respectively. The weighted column-grained representation V_{F} and the row-grained representation V_{C} are calculated in this way:

$$\begin{aligned} V_{F} &= M_{F} \odot f_{1}(Linear_{1}(H^{t})), \\ V_{C} &= M_{C} \odot f_{1}(Linear_{1}(H^{t})), \end{aligned} \tag{7}$$

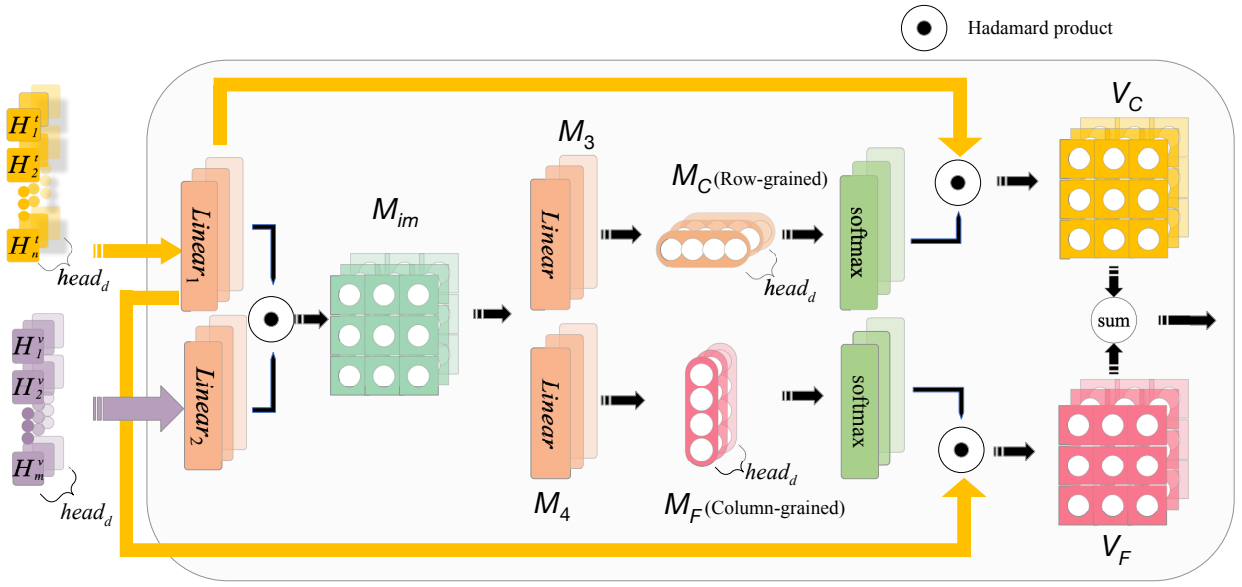


Figure 4: Structure of Multi-Granularity Excavator (MGE).

where $V_F, V_C \in R^{n \times d}$. Then V_F and V_C are added together to get the final output:

$$X^v = f_3(SUM(V_F, V_C)), \tag{8}$$

where $f_3(\cdot)$ denotes an encoder, the dimension of its output is the same as M_{im}. SUM denotes the sum operation. The calculation of MGE can be simplified to the following form:

$$\begin{aligned} X^v &= MGE(H^v, H^t), \\ T^t &= MGE(H^t, X^v), \end{aligned} \tag{9}$$

where X^v and T^t denotes the output of image-guided and text-guided MGE, respectively. In this way we consider both row-grained and column-grained fusion to get the final fused representation.

3.3.2. Interweave Encoder

The Interweave Encoder consists of two MGE, more than one can be stacked and each layer of Interweave Encoder shares the same encoder. After experiments, we find that one Interweave Encoder is the best choice. The input of the MGE consists of the outputs of the MGE at different interweaving degrees. For the lth Interweave Encoder, we first use image-guided MGE:

$$X_l^v = X_{l-1}^v + MGE(X_{l-1}^v, T_{l-1}^t), \tag{10}$$

where X_l^v denotes the output of the image-guided MGE in the lth Interweave Encoder, T_l^t denotes the output of the text-guided MGE in the lth Interweave Encoder. Text-guided MGE:

$$T_l^t = T_{l-1}^t + MGE(T_{l-1}^t, X_l^v). \tag{11}$$

It is worth noting that each layer of our text or image encoder shares the same encoders, respectively, enabling it to encode multi-granularity features that contain richer information.

3.3.3. Pre-Post Adaptive Gate

In the presence of image-text mismatch, integrating mismatched images will introduce irrelevant noise, leading to the textual features fused with image carrying a significant amount of noise that severe impact classification tasks. To address this issue, we introduce a learnable parameter I within $[0, 1]$ to adjusts the proportion of fused features and

textual features when combining, enabling the suppression of noise propagation. Similar to but different from [1], I is defined as follow:

$$I = \sigma(W_1 T^t + W_2 H^t + b), \tag{12}$$

where $W_1, W_2 \in R^{d\times d}, b \in R^{d\times 1}, T^t = \{T_1^t, T_2^t, ..., T_n^t\}$ denotes the post-fusion textual features and $H^t = \{H_1^t, H_2^t, ..., H_n^t\}$ denotes the pre-fusion textual features, and the final features input into the CRF are calculated as follows:

$$z = I \odot H^t + (1 - I) \odot T^t. \tag{13}$$

The more mismatched between image and text are, the more noise will be carried in T^t, the $1 - I$ will be more closer to 0, so that the information comes from T^t will be less.

3.4. Implicit Positive Sample Contrastive Learning

Our second innovation is Implicit Positive Sample Contrastive Learning (IPS-CL). In this section we will introduce the Implicit Positive Sample Mining Strategy and IPS-CL calculation method.

3.4.1. Implicit Positive Sample Mining Strategy

It is widely recognized that contrastive learning can efficiently minify the semantic distance between positive samples, while push away the semantic distance between negative samples, making a strategy for forming positive and negative samples a crucial factor in determining the performance of contrastive learning.

According to [30], in more than 33.8% of tweets the image does not relate to the text at all, which indirectly indicates that in more tweets, the image does not relate to the text completely. And we find that images in other tweets often have potential relevance to the text. These images can provide effective semantic information that the original image cannot provide, and can play an important role in MNER. Standard contrastive learning takes the original image-text pairs in the same batch as the positive sample pairs without considering the above problems. Therefore, different from the existing standard contrastive learning, we select samples with high potential relevance between different text-image pairs as Implicit Positive Sample Pairs.

As in Fig.5, each point denotes the size of potential relevance coefficient between text features and visual features in the same batch, the larger the point indicates the higher the potential relevance coefficient, and our Implicit Positive Sample Mining Strategy is to select K images for text to form Implicit Positive Sample Pairs based on their potential relevance.

3.4.2. Implicit Positive Sample Contrastive Learning

Based on the above strategy, we introduce a contrastive learning approach that aims to select the top K Implicit Positive Sample Pairs with the highest relevance. Specifically, given a textual feature $t_i \in R^{n\times d}$, compute the potential relevance coefficient between visual features $V = \{v_1, v_2, ..., v_N\}$ and t_i in the same batch, where n denotes the batch size, $v_j \in R^{m\times d}$. Then we compute the potential relevance coefficient between them:

$$sim(v_j, t_i) = \frac{f(v_j)^T g(t_i)}{\| f(v_j) \| \| g(t_i) \|}, \tag{14}$$

where $f(\cdot)$, $g(\cdot)$ denotes two encoders, $sim(v_j, t_i)$ denotes the relevance between the image v_j and the text t_i, $sim(V, T) \in R^{N\times N}, T = \{t_1, t_2, ..., t_N\}$. We select K images with the highest relevance to t_i:

$$V_K = \{v_{K_1}, v_{K_2}, ..., v_{K_K}\}. \tag{15}$$

IPS-CL are calculated according to the following formula:

$$Loss_{IPS-CL} = \frac{1}{N} \sum_{i=1}^{N} -log \frac{\sum_{m=1}^{K} e^{sim(t_i, v_{K_m})/\tau}}{\sum_{j=1}^{N} e^{sim(t_i, v_j)/\tau}}, \tag{16}$$

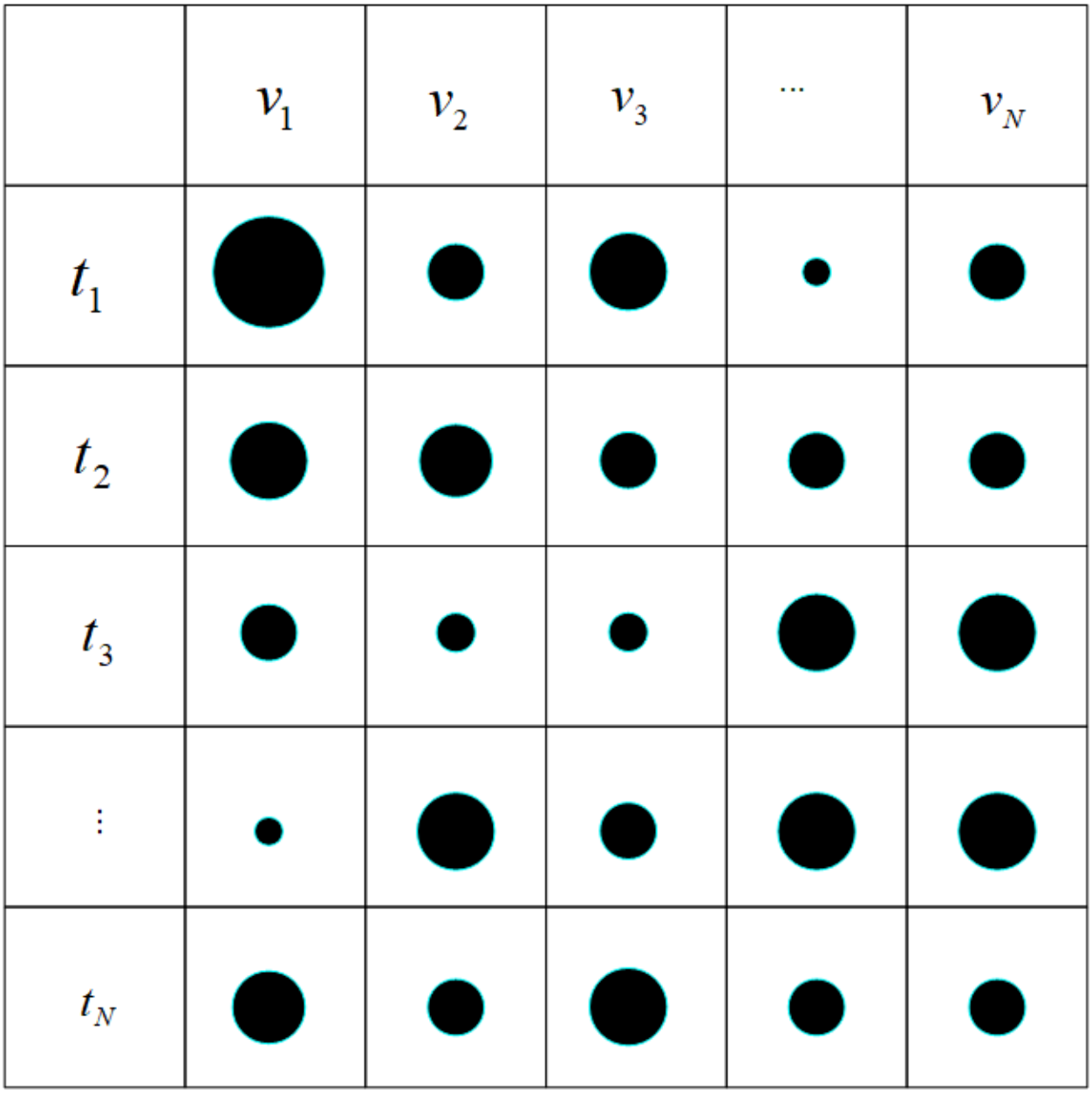


Figure 5: Explanation of the Implicit Positive Sample Mining Strategy. The size of the dot represents the potential relevance coefficient (eq.(14)).

where τ denotes the temperature coefficient and N denotes the batch size.

As only in 33.8% of the tweets [30], the image does not reflect the text at all. It means that in the remaining 66.2% of the tweets, there is a certain relevance between text-image pairs. So standard contrastive learning can indeed play a certain role. The standard contrastive loss is defined as follows:

$$Loss_{CL} = \frac{1}{N}\sum_{i=1}^{N} -log\frac{e^{sim(t_i,v_i)/\tau}}{\sum_{j=1}^{N} e^{sim(t_i,v_j)/\tau}}. \tag{17}$$

Table 1
Two multimodal twitter datasets in a statistical summary.

Entity type	Twitter-2015				Twitter-2017			
	Train	Dev	Test	Total	Train	Dev	Test	Total
PER	2,217	552	1,816	4,585	2,943	626	621	4,190
LOC	2,091	552	1,697	4,340	731	173	178	1,082
ORG	928	247	839	2,014	1,674	375	395	2,444
MISC	940	225	726	1,891	701	150	157	1,008
Total Entity	6,176	1,546	5,078	12,800	6,049	1,324	1,351	8,724
Num of Tweets	4,000	1,000	3,257	8,257	3,373	723	723	4,819

3.5. MNER Decoder

Conditional Random Fields (CRFs) show a strong capacity to obtain information from semantic space for the labeling of sequence in MNER [46, 39, 45, 6, 49, 36]. So we consider CRF as our MNER decoder and calculates the prediction $Loss_{MNER}$ simultaneously with the Implicit Positive Sample Contrastive Loss and the Standard Contrastive Loss. So $Loss_{MNER}$ is define as follows:

$$Loss_{MNER} = -\frac{1}{|D|}\sum_{j=1}^{N}(logP(y_j|S_j, V_j)), \tag{18}$$

where $D = \{S_j, V_j, y_j\}_{j=1}^{N}$ denote the traning examples. So the final loss of MGIPS-CL is be defined as follows:

$$Loss = Loss_{MNER} + Loss_{IPS-CL} + Loss_{CL}. \tag{19}$$

4. Experiment

4.1. Datasets

We use two public datasets collected from Twitter, twitter-2015 [46] and twitter-2017 [22]. The label schema is BIO [25], The statistical summary of two twitter datasets is shown in Table 1. For the missing images, we use the default image as a substitute like [46].

4.2. Baselines

With our approach, we focus on the comparison of three groups of baseline methods.

Several representative text-based NER methods are included in the first group: BiLSTM-CRF [16], CNN-BiLSTM-CRF [24], BERT [13], BERT-CRF and BERT-BiLSTM-CRF.

The second group contains several representative LLM MNER models: ChatGPT, GPT-4, CoT (ChatGPT) [3] and UMIE (Flan-T5) [28].

The third group contains several representative Multimodal MNER models: AdaCAN [46], AdaCAN-BERT-CRF, UMT [43], MEGA [50], UMGF [45], MAF [41], M3S [31], CAT-MNER [35], FMIT [23], HVPNet [7], MKGFormer [6], BFCL [32], MNER-QG [12], DebiasCL [47], MPMRC-MNER [2], BGA-MNER [4], MCG-MNER [37], TMR [49], ICKA [44], VEC-MNER [36].

4.3. Details

We use $CLIP_{IMAGE}$ and $BERT_{TEXT}$ to perform mutual-attention on text and visual embeddings. The dimension of hidden layer is set to 768. We set the maximum length of the sentence input as 40. We set the attention heads and layers of mutual-attention to 12 and 12, respectively. We use the warm-up training strategy. The warmup_ratio, lr of $BERT_{TEXT}$, learning rate of $CLIP_{IMAGE}$ and weight_decay are respectively set to 0.06, 5e-5, 3e-5 and 1e-2. The number of Implicit Positive Sample Pairs is set to 3. All experiments are performed on Python 3.8.18 and NVIDIA GTX 3090 GPUs with PyTorch 2.1.0.

Table 2
Performance Comparison on Two TWITTER Datasets. Results of all the models are the average of random five times.

Methods	Twitter-2015							Twitter-2017						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
Text Baselines														
BiLSTM-CRF (2016)	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.5	52.56	79.42	73.43	76.31
CNN-BiLSTM-CRF (2016)	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
BERT-BiLSTM-CRF (2019)	84.32	79.31	61.66	37.53	71.03	73.57	72.27	90.29	84.55	80.97	64.85	83.20	84.68	83.93
LLM Baselines														
ChatGPT	-	-	-	-	-	-	50.21	-	-	-	-	-	-	57.50
GPT-4	-	-	-	-	-	-	57.98	-	-	-	-	-	-	66.61
CoT (ChatGPT) (2023)	-	-	-	-	-	-	76.53	-	-	-	-	-	-	87.79
UMIE (Flan-T5) (2024)	-	-	-	-	-	-	76.10	-	-	-	-	-	-	88.10
Multimodal Baselines														
AdaCAN (2018)	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
AdaCAN-BERT-CRF (2018)	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
UMT (2020)	85.24	81.58	63.03	39.45	71.84	74.61	73.20	91.56	84.73	82.24	70.10	85.08	85.27	85.18
MEGA (2021)	-	-	-	-	70.35	74.58	72.35	-	-	-	-	84.03	84.75	84.39
UMGF (2021)	84.26	83.17	62.45	42.42	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51
MAF (2022)	84.67	81.18	63.35	41.82	71.86	75.10	73.42	91.51	85.80	85.10	68.79	86.13	86.38	86.25
M3S (2022)	86.05	81.32	62.97	41.36	74.92	75.14	75.03	92.73	84.81	82.49	69.53	86.93	85.21	86.06
CAT-MNER (2022)	85.57	81.97	61.12	40.20	76.19	74.65	75.41	91.90	85.96	83.38	68.67	87.04	84.97	85.99
FMIT (2022)	86.77	83.93	64.88	42.97	75.11	77.43	76.25	93.14	86.52	83.93	70.90	87.51	86.08	86.79
HVPNet (2022)	-	-	-	-	73.87	76.82	75.32	-	-	-	-	85.84	87.93	86.87
MKGFormer (2022)	-	-	-	-	-	-	-	-	-	-	-	86.98	88.01	87.49
BFCL (2023)	85.60	81.77	63.81	40.30	74.02	75.07	74.54	91.17	86.43	83.97	66.67	85.99	85.42	85.70
MNER-QG (2023)	85.31	81.65	63.41	41.32	77.43	72.15	74.70	90.90	86.19	84.52	71.67	88.26	85.65	86.94
DebiasCL (2023)	85.97	81.84	64.02	43.38	74.45	76.13	75.28	93.46	84.15	84.42	67.88	87.59	86.11	86.84
MPMRC-MNER (2023)	85.88	83.06	**66.60**	42.99	77.15	75.39	76.26	92.80	86.80	83.40	72.79	87.10	87.16	87.13
BGA-MNER (2023)	86.80	**83.62**	63.60	42.65	**78.60**	74.16	76.31	93.71	85.55	85.71	71.05	87.71	87.71	87.71
MCG-MNER (2023)	-	-	-	-	76.00	77.32	76.65	-	-	-	-	87.59	88.52	88.05
TMR (2023)	-	-	-	-	75.26	76.49	75.87	-	-	-	-	88.12	88.38	88.25
ICKA (2024)	87.01	83.85	65.87	48.28	72.36	**78.75**	75.42	93.99	**87.24**	**86.24**	**75.76**	85.13	**89.19**	87.12
VEC-MNER (2024)	86.11	81.03	62.86	40.60	74.56	75.23	74.89	93.88	81.27	85.49	73.40	87.42	87.61	87.51
MGIPS-CL (Ours)	**86.83**	82.72	66.20	**46.02**	77.70	76.04	**76.86**	**94.87**	85.88	85.68	73.31	**88.29**	88.75	**88.52**

4.4. Results

Analysis of unimodel methods. Among the unimodal models, the models using BERT outperform the models using CNN and LSTM in terms of Pre, Rec, and F1-score. This demonstrate the superiority of BERT as an encoder in NER. The models using CRF are all superior to the models without using CRF, which shows the superiority of CRF as a decoder in NER.

Analysis between unimodel and multimodel methods. In comparison to the unimodal model, the multimodal model exhibits clear superiority in MNER, which illustrates the effectiveness of images as auxiliary information in MNER. But the F1-score of AdaCAN-BERT-CRF on twitter-2015 is 0.12% lower than BERT-BiLSTM-CRF. At the same time, the F1-score of AdaCAN-BERT-CRF on the twitter-2017 dataset is 0.17% higher than the F1-score of BERT-BiLSTM-CRF. We believe that this is because the correlation between text-image pairs in twitter-2015 is low, while AdaCAN-BERT-CRF doesn't solve the problem of mismatch between the image and text. Our model achieves comprehensive SOTA on both twitter-2015 and twitter-2017 datasets, which indicates that our model can make full use of its effective semantic information while effectively filtering the noise brought by images.

Analysis of Large Language Model methods. Among the large model methods, the F1-score of CoT surpasses that of ChatGPT by 16.32% and 30.29% and exceeds GPT-4 by 18.55% and 21.18% on the Twitter-2015 and Twitter-2017 datasets, respectively. This indicates that CoT enhances the model's logical reasoning ability, leading to improved recognition performance. The F1-score of UMIE is 0.43% lower than that of CoT on Twitter-2015 but 0.31% higher on Twitter-2017. The reasons for this are as follows: (1) The proportion of image-text matching in the Twitter-2015 dataset is lower than in the Twitter-2017 dataset. (2) On the Twitter-2015 dataset, UMIE models more image-text correlations, introducing some noise, whereas CoT focuses more on logical reasoning.

Analysis of compared with all other MNER methods. On the Twitter-2015 dataset, our model improves F1-scores by 1.97%, 1.44%, 0.55%, 0.60% and 0.61% compared to VEC-MNER, ICKA, BGA-MNER, MPMRC-MNER and FMIT, respectively. On the Twitter-2017 dataset, our model improves F1-scores by 0.98%, 1.40%, 0.47%, 0.81%

Table 3
Ablation Study of MGIPS-CL.

Models	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
MGIPS-CL	**77.70**	**76.04**	**76.86**	**88.29**	**88.75**	**88.52**
MGIPS-CL w/o IPS-CL	74.66	77.08	75.85	87.20	88.23	87.71
MGIPS-CL w/o MGI	75.09	76.81	75.94	86.87	87.92	87.39
MGIPS-CL w/o Row-grained	76.03	76.70	76.36	88.25	87.92	88.08
MGIPS-CL w/o Column-grained	76.12	76.70	76.41	87.79	88.52	88.15

and 1.03% compared to VEC-MNER, ICKA, MCG-MNER, BGA-MNER and MKGFormer, respectively. Our model achieves the SOTA performance of 75.86% and 88.52% on both datasets, with an F1-score improvement of 0.21% over the previous best MCG-MNER on Twitter-2015 and 0.27% over the previous best TMR on Twitter-2017. This demonstrates that: (1) The proposed Multi-Granularity Interweave module effectively integrates visual and textual features in a multi-granularity manner, achieving deep fusion of the two modalities. (2) The proposed Implicit Positive Sample Contrastive Learning leverages the latent correlations between text and images to mine Implicit Positive Sample Pairs, which can avoid the visual noise introduced by image-text mismatch and fully mine the visual information that is truly associated with the text.

4.5. Ablation

In this section, we further analyze the effectiveness of our model by conducting experiments on variants of our model. we perform a comparison between the MGIPS-CL and its ablations concerning the Implicit Positive Sample Contrastive Learning **(w/o IPS-CL)**, the Multi-Granularity Interweave Module**(w/o MGI)**, the Row-grained**(w/o Row-grained)** and the Column-grained**(w/o Column-grained)**. Table 3 shows the experimental results of our model after ablating different modules.

Importance of IPS-CL Module. The F1-score of MGIPS-CL **w/o IPS-CL** drop 1.01% and 0.81% on the twitter-2015 and twitter-2017 datasets, respectively. For the reason of the severe performance degradation on the twitter-2015 dataset, we believe that it is mainly due to after losing the assistance of the K potential related images, the shortcoming of insufficient semantic information in the image is seriously exposed, resulting in a sharply drop in performance after **w/o IPS-CL**. In conclusion, the results on both datasets show that our IPS-CL plays an important role in our model.

Importance of MGI Module. The F1-score of MGIPS-CL **w/o MGI** drop 0.92% and 1.13% on the twitter-2015 and twitter-2017 datasets, respectively. This is because after removing MGI, our model loses the ability to interweave different modal information to mine rich semantics and filter out noise, leading to a decrease in the performance of the model. This result fully demonstrates the importance of our MGI.

Importance of Multi-Granularity. The F1-score of MGIPS-CL **w/o Row-grained** drop 0.50% and 0.44% on the twitter-2015 and twitter-2017 datasets, respectively. Then the F1-score of MGIPS-CL **w/o Column-grained** drop 0.45% and 0.37% on the two datasets, respectively. However, these F1-scores for single-granularity fusion are all higher than those **w/o MGI** and much lower than MGIPS-CL. The result further confirms the effectiveness of our MGI for deep fusion of images and texts at a multi-granularity way. These results strongly illustrate the important role of our MGI in improving model performance.

4.6. Parameter sensitivity analysis

4.6.1. K-value sensitivity analysis

Section 3 mentions that we select the K text-image pairs with the highest relevance as Implicit Positive Sample Pairs for IPS-CL. The value of K seriously affects the amount of Implicit Positive Sample Pairs, so it is worthwhile to pay attention to the value of K.

We set the value of K=[1, 2, 3, 4, 5, 6], and train 40 epochs with the same hyperparameters. The experimental results are shown in Fig.6(a) and Fig.6(b), which show that the model's effect is poor when K is small. The model performs best when K=3, and the model's performance is poorer with the increase of K when K is larger than 3. We believe that the reason is when the value of K is small, the number of Implicit Positive Sample Pairs (IPSP) is too

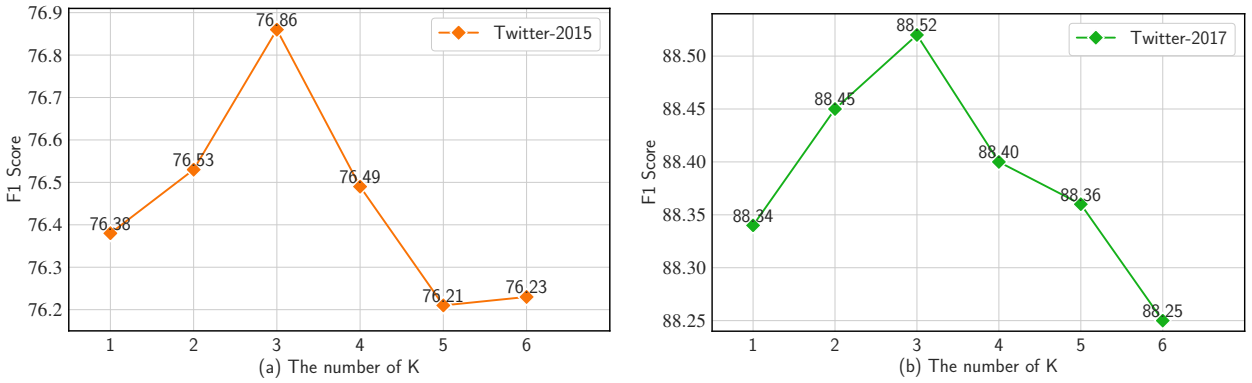


Figure 6: Parameter sensitivity analysis experiment for the number of Implicit Positive Sample Pairs.

small, so that the role of IPSP is not fully utilized. When K is larger than 3 and gradually larger, the number of IPSP are gradually increased, leading to the selection of more and more incorrect IPSP. The above results show that an appropriate value of K can help the model effectively mine IPSP to supplement the missing information in the original image.

4.6.2. Interweave Encoder Quantitative Sensitivity Analysis

Stacking different numbers of Interweave Encoder shown in Fig.3 will affect the degree of interweaving, so it is worthwhile to pay attention to the number of Interweave Encoder. We set the number of Interweave Encoder to [1, 2, 3, 4, 5, 6, 7] and conduct experimental analysis.

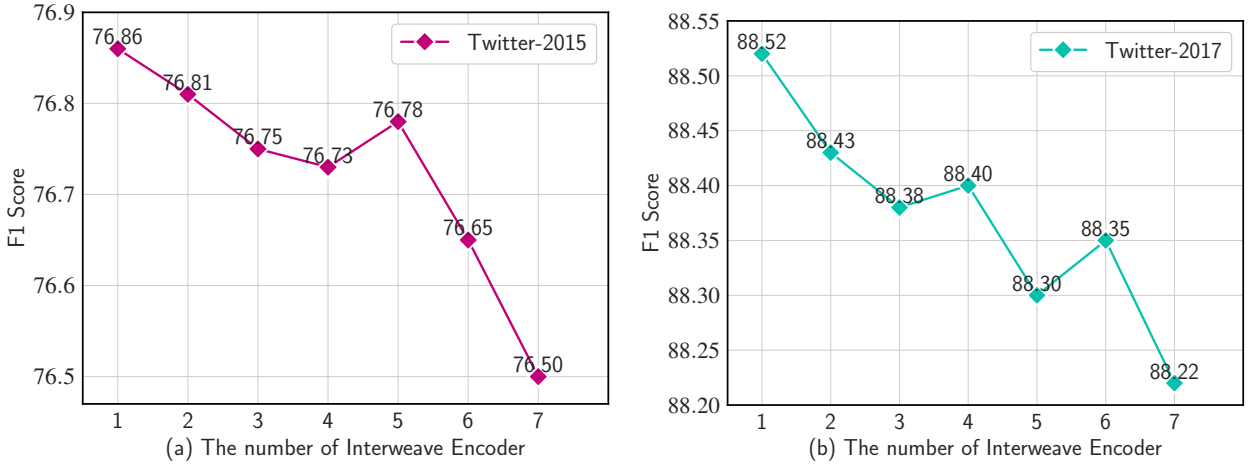


Figure 7: Parameter sensitivity analysis experiment for the number of Interweave Encoder.

Fig.7 shows that the F1-score on twitter-2015 and twitter-2017 datasets reaches the highest when there is only one Interweave Encoder. There is no clear relationship between F1-score and the quantity of Interweave Encoder. We believe it is because our Interweave Encoder has excellent performance so that using only one Interweave Encoder is already good enough to fuse visual textual features effectively. Stacking more than one Interweave Encoder will lead to the problem of over-interweaving. While it will fuse effective information and filter the noise to improve the model accuracy, it will also inevitably cause the loss of effective information and reduce the generalization ability of the model. The different degrees of improvement and degradation lead to the phenomenon that the F1-scores of different numbers of Interweave Encoder appear to be alternately high and low, with no obvious pattern.

4.7. Low-Resource Scenario Analysis

Owing to the absence of publicly accessible code for BFCL, we are constrained from performing experiments with it. Consequently, we choose TMR and VEC-MNER to serve as the comparative baseline for the experiment.

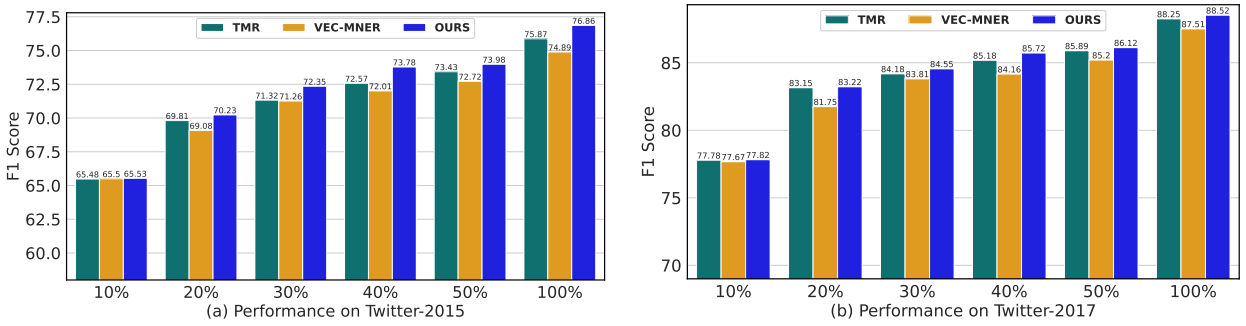


Figure 8: F1-scores of Low-Resource on two datasets.

Same as [26], we aim to investigate the efficacy of MGIPS-CL under low-resource conditions by carrying out training on the Twitter-2015 and Twitter-2017 datasets at varying scales. These scales encompass 10%, 20%, 30%, 40%, 50%, and the full 100% of the available training data. As shown in Fig.8, the experimental results show that our model outperforms TMR and VEC-MNER even at a small training scale. Especially on the Twitter-2017 dataset, MGIPS-CL achieves the performance of BFCL at 100% scale at only 50% training scale. And as the scale of the data increases, the advantages of our model gradually increase. Our model only outperforms VEC-MNER on two datasets when the data scale is large. This is because as the data scale increases, the number of implicit positive sample pairs will also increase, which can fully utilize the performance of our model. This indicates that our model has higher performance on larger datasets.

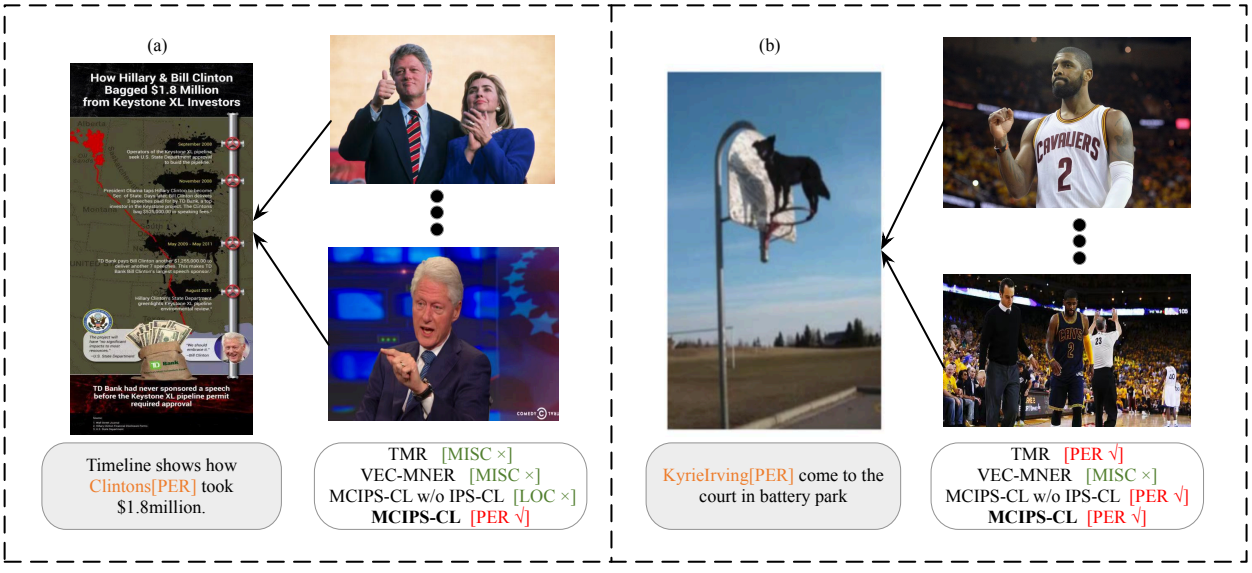
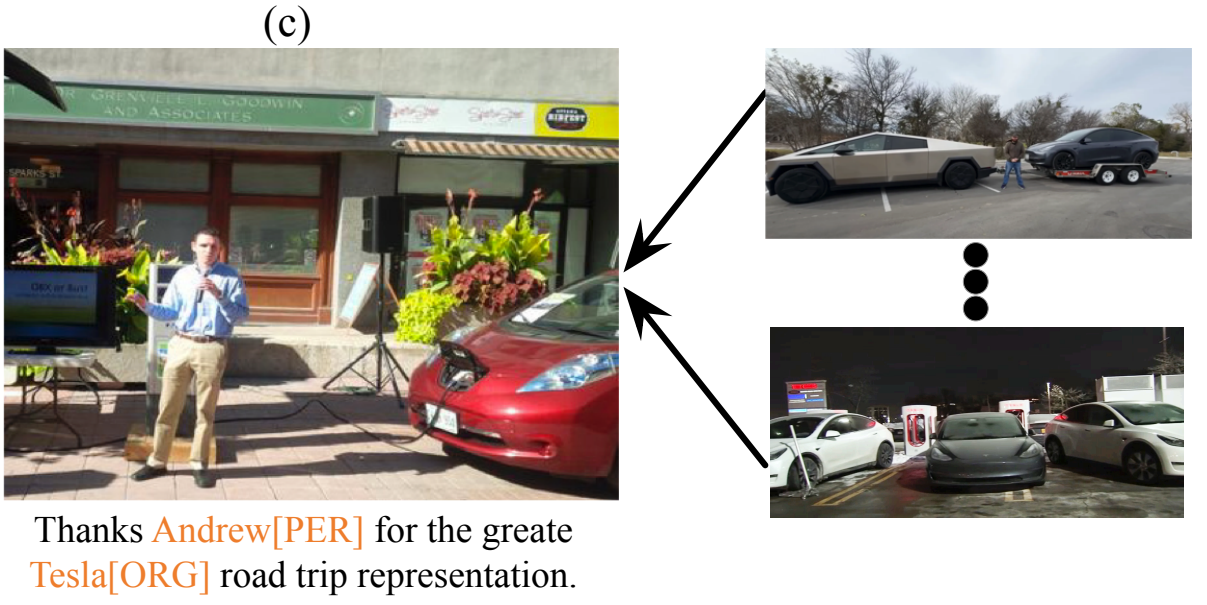


Figure 9: The results of MGIPS-CL compared with TMR, VER-MNER and MGIPS-CL **w/o IPS-CL**.

4.8. Case Study

To demonstrate the power of our model in a more intuitive way, as shown in Fig.9 and Fig.10, we select three typical extraction cases.



UMT:	1-PER	√	2-None	×
BFCL:	1-PER	√	2-PER	×
TMR:	1-PER	√	2-MISC	×
VEC-MNER:	1-PER	√	2-ORG	√
MCIPS-CL:	1-PER	√	2-MISC	×

Figure 10: The results of MGIPS-CL compared with UMT, BFCL, TMR and VEC-MNER.

In case (a), we observe objects like a bag of dollar bills, a map, and a pipeline in the image. Although there is some OCR text description related to the entity "Clintons" in the image, the ResNet and CLIP image encoders primarily focus on understanding the object information in the image and have limited capability to comprehend OCR text. The presence of a large amount of irrelevant information misleads the model, causing TMR, VER-MNER, and MGIPS-CL **w/o IPS-CL** to incorrectly classify "Clintons" as "MISC" or "LOC". However, after incorporating IPS-CL, the model associates more relevant images to help accurately classify "Clintons" as "PER", demonstrating that IPS-CL can effectively select and utilize implicit positive sample pairs in social media scenarios.

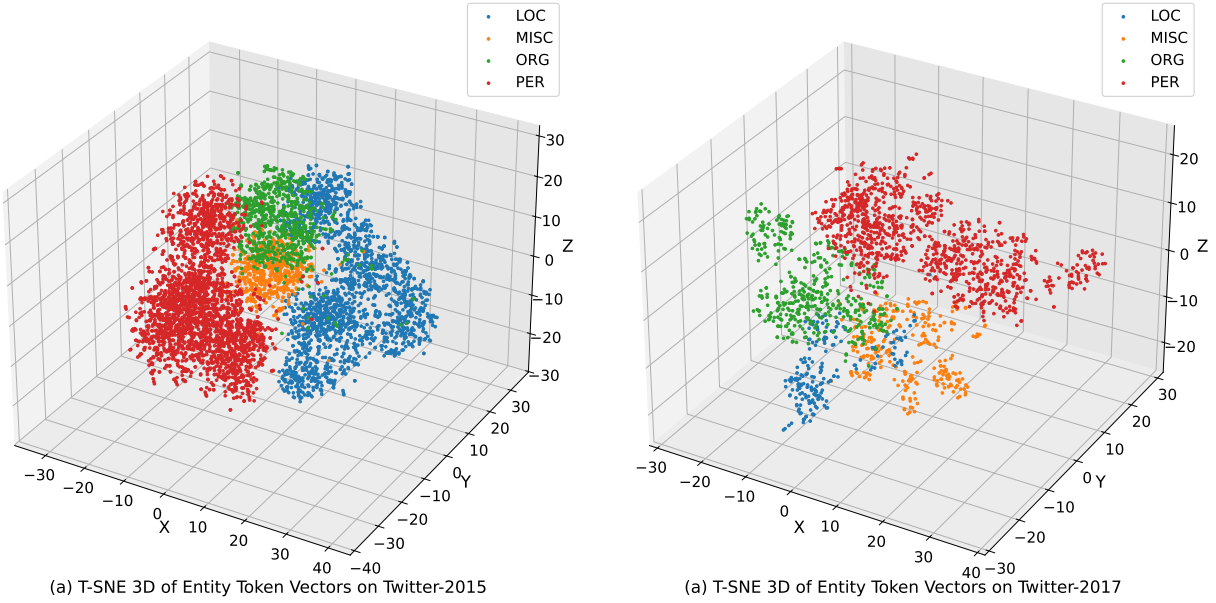
In case (b), the image depicts a dog standing on a basket, which is unrelated to the text. As a result, the visual information becomes noise, leading VEC-MNER to incorrectly classify "Kyrie Irving" as "MISC". However, due to the presence of the MGI module, both MGIPS-CL **w/o IPC-CL** and MGIPS-CL demonstrate strong noise filtering capabilities, correctly identifying "Kyrie Irving" as "PER". Notably, after incorporating IPS-CL, the model still accurately recognizes "Kyrie Irving" as "PER", indicating that the IPS-CL mechanism effectively maintains the model's ability to judge the semantic relevance between text-image pairs without excessively narrowing the semantic distance of irrelevant text-image pairs, thereby avoiding misclassification.

In case (c), the result shows the misrecognition problem of MGIPS-CL. UMT, BFCL and TMR believe that images are highly correlated with text [32], resulting in excessive introduction of image information, so incorrectly predict "Tesla" as "PER". VEC-MNER can mine the deep semantic association between visual objects and entities, so it correctly identified Tesla as ORG. However, our model incorrectly predict it as "MISC". Because MGIPS-CL select

the image of the Tesla car when selecting the Implicit Positive Sample Pairs, our model incorrectly identify "Tesla" as "MISC".

4.9. Visualization

To verify the effectiveness of the MGIPS-CL model on the multimodal named entity recognition (MNER) task, we conducted a 3D t-SNE visualization analysis of the entity label vectors on the Twitter-2015 and Twitter-2017 test datasets. T-SNE, a commonly used dimensionality reduction technique, projects high-dimensional vectors into a 3D space, enabling intuitive visualization of the distribution and clustering of different entity types (e.g., PER, LOC, ORG, MISC). This allows for an assessment of the model's capability in distinguishing entities and maintaining semantic consistency.



(a) T-SNE 3D of Entity Token Vectors on Twitter-2015 (a) T-SNE 3D of Entity Token Vectors on Twitter-2017

Figure 11: T-SNE 3D visualization of Entity Token Vectors on two test datasets.

As shown in Fig.11a, the 3D t-SNE visualization results on the Twitter-2015 test dataset demonstrate clear clustering patterns for different entity types. Specifically, vectors representing PER (person) entities form a compact cluster, indicating that the model effectively captures semantic features of personal names. LOC (location) and ORG (organization) entity vectors also exhibit high intra-class cohesion, suggesting strong discriminative capability for these categories. Despite the semantic diversity of MISC (miscellaneous) entities, their vectors still display a certain degree of aggregation, reflecting the model's ability to extract shared semantic features. Moreover, the distinct spatial separation among different entity types further validates the MGIPS-CL model's capacity to distinguish entities and reduce semantic confusion.

As shown in Fig.11b, the 3D t-SNE visualization on the Twitter-2017 test dataset reveals distinct clustering patterns across different entity types, consistent with observations on the Twitter-2015 dataset. Notably, PER (person) entity vectors exhibit even tighter clustering, indicating enhanced recognition capability for personal names. LOC (location) vectors demonstrate greater concentration, suggesting the model's improved ability to filter noise and extract salient features. ORG (organization) entities form well-defined clusters, reflecting high accuracy and consistency in organizational entity recognition. Although MISC (miscellaneous) vectors are more dispersed due to their inherent semantic diversity, they still display observable aggregation, implying that the model can capture their latent commonalities. Compared to the Twitter-2015 dataset, the entity vector distribution on Twitter-2017 is more distinct, with clearer inter-class separations, further validating the robustness and effectiveness of the MGIPS-CL model on more complex data.

Overall, the 3D t-SNE visualization results on both the Twitter-2015 and Twitter-2017 test datasets demonstrate that the MGIPS-CL model effectively distinguishes between different entity types, exhibiting clear clustering patterns. These results highlight the effectiveness of the model's key components: (1) The Multi-Granularity Interleaving (MGI) module, which enhances entity representation by fusing visual and textual features at multiple granularities while mitigating noise propagation; (2) The Implicit Positive Sample Contrastive Learning (IPS-CL) mechanism, which exploits latent correlations between text and images to improve the models semantic discrimination and understanding of entities.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose MGIPS-CL, a novel multimodal named entity recognition (MNER) model that achieves state-of-the-art performance on the Twitter-2015 and Twitter-2017 datasets, with F1 scores of 76.86% and 88.52%, respectively. The superior performance of MGIPS-CL can be attributed to three key innovations: (1) the Implicit Positive Sample Contrastive Learning (IPS-CL) mechanism, which effectively leverages latent correlations between text and image modalities to mitigate noise from mismatched pairs and enhance semantic understanding of entities; (2) the Multi-Granularity Interleaving (MGI) module, which enables deep cross-modal fusion by integrating visual and textual features at multiple granularities, thereby reducing noise propagation and enhancing key information extraction; and (3) the model's strong robustness and generalization capabilities, as demonstrated by its competitive performance even under low-resource conditions.

Despite its strong performance, MGIPS-CL remains sensitive to the manually predefined number of implicit positive pairs (K) in IPS-CL. This dependency limits its adaptability to varying data complexities and task scenarios. As future work, we aim to introduce a learnable and adaptive mechanism for determining K based on data semantics and context, enabling the model to dynamically select the optimal number of positive pairs. We also plan to explore richer contextual and semantic representations to further improve the models recognition accuracy and robustness. These enhancements are expected to expand the applicability of MGIPS-CL to a broader range of multimodal tasks and real-world settings.

Acknowledgments

This work is supported by the Key Program of the National Natural Science Foundation of China (Grant No.62237001).

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[2] Xigang Bao, Mengyuan Tian, Zhiyuan Zha, and Biao Qin. Mpmrc-mner: A unified mrc framework for multimodal named entity recognition based multimodal prompt. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 47–56, 2023.

[3] Feng Chen and Yujian Feng. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*, 2023.

[4] Feng Chen, Jiajia Liu, Kaixiang Ji, Wang Ren, Jian Wang, and Jingdong Chen. 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, 2023.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 904–915, 2022.

[7] Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 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, 2022.

[8] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, 2021.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[10] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

[11] Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. ICLR, April 2019.

[12] Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8032–8040, 2023.

[13] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[15] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13171–13179, 2021.

[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[18] Wei Li, Yajun Du, Xianyong Li, Xiaoliang Chen, Chunzhi Xie, Hui Li, and Xiaolei Li. Ud_bbc: Named entity recognition in social network combined bert-bilstm-crf with active learning. *Engineering Applications of Artificial Intelligence*, 116:105460, 2022.

[19] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, 2021.

[20] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, 2020.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[22] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 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, 2018.

[23] Junyu Lu, Dixiang Zhang, Jiaxing Zhang, and Pingjian Zhang. Flat multi-modal interaction transformer for named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2055–2064, 2022.

[24] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, 2016.

[25] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. In *Proceedings of NAACL-HLT*, pages 852–860, 2018.

[26] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845, 2023.

[27] 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.

[28] Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. Umie: Unified multimodal information extraction with instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19062–19070, 2024.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[30] Alakananda Vempala and Daniel Preoţiuc-Pietro. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840, 2019.

[31] Jie Wang, Yan Yang, Keyu Liu, Zhiping Zhu, and Xiaorong Liu. M3s: Scene graph driven multi-granularity multi-task learning for multi-modal ner. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:111–120, 2022.

[32] Peng Wang, Xiaohang Chen, Ziyu Shang, and Wenjun Ke. Multimodal named entity recognition with bottleneck fusion and contrastive learning. *IEICE TRANSACTIONS on Information and Systems*, 106(4):545–555, 2023.

[33] Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. Ita: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, 2022.

[34] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 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, 2022.

[35] Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022.

[36] Pengfei Wei, Hongjun Ouyang, Qintai Hu, Bi Zeng, Guang Feng, and Qingpeng Wen. Vec-mner: Hybrid transformer with visual-enhanced cross-modal multi-level interaction for multimodal ner. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 469–477, 2024.

[37] Junjie Wu, Chen Gong, Ziqiang Cao, and Guohong Fu. Mcg-mner: A multi-granularity cross-modality generative framework for multimodal ner with instruction. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3209–3218, 2023.

[38] Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. Denoising bottleneck with mutual information maximization for video multimodal fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2231–2243, 2023.

[39] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 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, 2020.

[40] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*, 2020.

[41] Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 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, 2022.

[42] Jun Yang, Liguo Yao, Taihua Zhang, Chieh-Yuan Tsai, Yao Lu, and Mingming Shen. Integrating prompt techniques and multi-similarity matching for named entity recognition in low-resource settings. *Engineering Applications of Artificial Intelligence*, 144:110149, 2025.

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

[44] Qingyang Zeng, Minghui Yuan, Jing Wan, Kunfeng Wang, Nannan Shi, Qianzi Che, and Bin Liu. Icka: an instruction construction and knowledge alignment framework for multimodal named entity recognition. *Expert Systems with Applications*, 255:124867, 2024.

[45] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 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, 2021.

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

[47] Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 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, 2023.

[48] Gang Zhao, Guanting Dong, Yidong Shi, Haolong Yan, Weiran Xu, and Si Li. Entity-level interaction via heterogeneous graph for multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6345–6350, 2022.

[49] Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei, and Qing Li. Rethinking multimodal entity and relation extraction from a translation point of view. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6810–6824, 2023.

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