IQCR-MI: Image Quality and Cross-Modal Relevance Guided Multi-Image Integration for Multimodal Named Entity Recognition

Abstract. Multimodal Named Entity Recognition (MNER) is an important research direction in the field of Natural Language Processing (NLP), which aims to enhance the performance of named entity recognition (NER) via additional image information. However, as posts with multiple images on social media are increasing, existing methods primarily focus on single-image scenarios, and there is still a significant gap in research on multi-image MNER scenarios. Existing multi-image approaches often treat all images uniformly, which may introduce the noise and even lead to biased model predictions via blindly aggregating low-quality and irrelevant images. To overcome this limitation, we propose an novel multi-image weighted integration framework (IQCR-MI) for the MNER-MI task, which dynamically adjusts the contribution of different images via introducing image quality and cross-modal relevance to guide multi-image feature fusion. Experiments show that our method performs better than existing methods in multi-image MNER tasks, providing a novel solution for the MNER-MI task.

Keywords: Multimodal Named Entity Recognition · Multi-image Cross-modal Correlation · Image Quality Assessment.

1 Introduction

Named entity recognition (NER), as a fundamental task in natural language processing, focuses on identifying predefined types of entities from the given text and serves as a crucial component in information extraction systems. With the increasing prevalence of multimodal content on social media platforms like Twitter, Instagram, and Facebook, multimodal named entity recognition (MNER) has gained more attention, which aims to enhance the performance of the NER model via integrating the visual modality as additional auxiliary information. Nevertheless, current MNER studies [23, 18, 29, 27] mainly concentrate on single-image scenarios, despite empirical analysis [33] showing that over 42% of real-world tweets contain multiple images. To address this challenge, Huang et al. [7] formally defined the multi-image MNER task and constructed the MNER-MI dataset as its benchmark. Meanwhile, they developed a temporal prompt model for multi-image scenarios (TPM-MI) that conceptualizes multiple images as sequential video frames, leveraging them as visual prompts to strengthen multimodal interaction.

Existing methods [7] for MNER-MI typically rely on an equalized aggregation strategy for multi-image features, which implicitly assumes that all visual

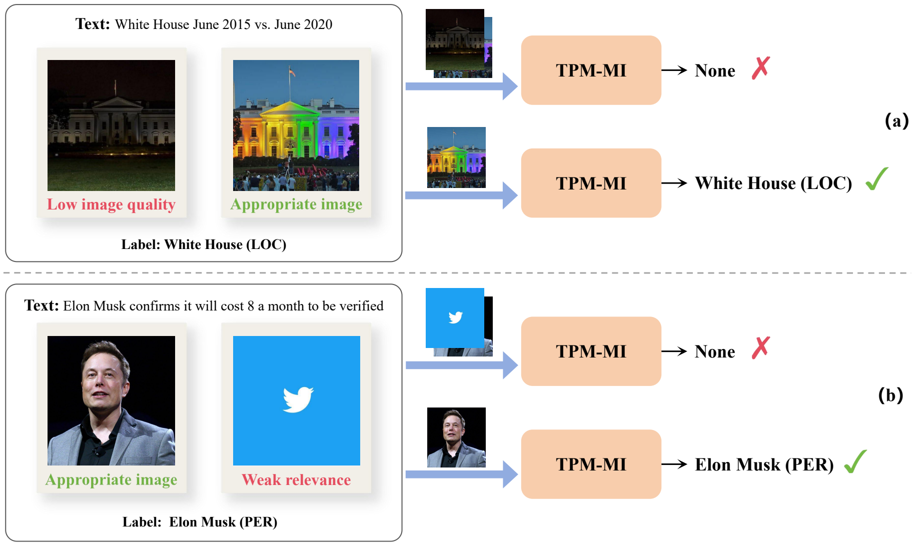

Fig. 1. Two illustrative cases using TPM-MI, the state-of-the-art method for the MNER-MI task, demonstrate our motivation. Figure (a) highlights how image quality affects entity recognition, while Figure (b) illustrates the impact of text-image relevance.

information contributes positively to entity recognition. However, in real-world scenarios, multiple images associated with the same tweet often exhibit substantial variance in quality metrics (e.g., brightness, clearness) and varying degrees of contextual relevance. We believe that low-quality and irrelevant images may inherently act as noise, and the current average feature aggregation strategy for multi-image could amplify such noise thereby limiting the model's performance improvement. Figure 1(a) and (b) respectively illustrate the impact of variations in image quality and contextual relevance on the performance of TPM-MI [7], the state-of-the-art method for the MNER-MI task. Figure 1(a) presents images of the White House taken at night in 2015 and 2020. We observe that the TPM-MI model accurately identifies the 'White House' entity when only the high-quality 2020 image is provided. However, its performance significantly deteriorates when both the 2015 and 2020 images are input simultaneously. This decline may be attributed to the poor quality of the 2015 image: insufficient lighting conditions hinder the extraction of effective features. When these low-quality features are combined with the high-quality ones from the 2020 image, substantial noise is introduced, which interferes with the model's judgment and ultimately leads to prediction errors. Then again, in the case of Figure 1(b), we need to identify the specific entity "Musk" in the text based on the multimodal context. Through comparative analysis, we can observed that the TPM-MI model's prediction errors stem from its improper integration of the Twitter icon image that are irrelevant to the context. Therefore, we believe that incorporating such low-quality or irrelevant images could introduce noise and potentially induce cognitive biases in the model's prediction, thereby leading to a performance bottleneck.

To this end, we propose an novel multi-image integration framework (IQCR-MI) to tackle these challenges, which adjusts the contribution of each image in the MNER-MI task based on image quality and cross-modal relevance. In detail, our approach leverages the state-of-the-art image quality assessment method to evaluate the quality of images. Additionally, we utilize CLIP to measure the relevance between multiple images and the corresponding text. Our approach assigns higher weights to highly relevant images while minimizing the negative influence of low-quality or irrelevant ones, which can help mitigate the performance bottlenecks caused by those noisy image data.

- We find that, in real-world scenarios, multiple images corresponding to the same tweet often show significant differences in quality metrics, such as brightness and clarity, as well as varying levels of contextual relevance. Existing MNER-MI methods typically aggregate each image uniformly, which significantly deviates from real-world scenarios.
- We design an novel multi-image integration framework (IQCR-MI) based on image quality and cross-modal relevance to address the challenges in multi-image MNER tasks. This framework dynamically adjusts the contribution of each image, significantly improving the overall performance of the MNER task.
- We conducted extensive experiments on the MNER-MI and MNER-UNI datasets. The results show that our approach (IQCR-MI) outperforms existing MNER-MI approaches, confirming the effectiveness of the multi-image integration mechanism in enhancing entity recognition performance in multi-image scenarios.

2 Related Work

2.1 Multimodal Named Entity Recognition

With growing abundance of multimodal data shared on social media, MNER has gradually gained more attention. MNER aims to improve the performance of NER via incorporating multimodal information, such as images.

Research on MNER could be categorized into three lines: methods based on text and image representation, knowledge-enhanced methods, and methods based on cross-modal attention. Methods based on text and image representation aim to effectively integrate text and image information [17, 33, 24]. For example, Wang et al. [24] introduced complex text encoders to enhance the semantic representation of text. Knowledge-enhanced methods improve MNER model performance by incorporating additional knowledge [18, 27, 25, 23]. For example, Ok et al. [18] enhance MNER by extracting external knowledge from both text and images. Methods based on cross-modal attention focus on using attention mechanisms to implicitly align and fuse the semantic information of text and image modalities [29, 34]. For example, Yu et al. [29] proposed a unified multimodal transformer architecture, which effectively promotes cross-modal interaction.

In real social media, such as twitter, a tweet often contains multiple images. However, the approaches mentioned above mainly focus on single-image scenarios, which creates a gap with the real-word application. To this end, Huang et al. [7] proposed the multi-image MNER task and developed a corresponding dataset, MNER-MI. Furthermore, they also introduced a multi-image temporal prompt model (TPM-MI), which treats multiple images as video frames and utilizes them as prompts to enhance interaction with the text. However, they typically treate all images equally and assume that each one has a positive contribution for entity recognition, overlooking the negative impact of low-quality and irrelevant images on the model.

2.2 Image Quality Assessment

Quality Assessment (IQA) is a challenging task aimed at automatically predicting the perceptual quality of distorted images. As a crucial component of low-level computer vision, IQA plays a vital role in a wide range of applications. Recent research on the IQA task is divided into two categories: full-Reference image quality assessment (FR-IQA) and no-reference image quality assessment (NR-IQA). FR-IQA methods require a high-quality reference image as a baseline to compare the differences between the image to be assessed and the reference image. In contrast, NR-IQA methods do not rely on a reference image, which evaluates image quality based on the intrinsic features of the image itself. Early NR-IQA methods primarily relied on the statistical or local features of the images [4, 5, 16]. The introduction of multi-dimensional attention network (MAN) for NR-IQA [28] marks a breakthrough in this field. MANIQA not only captures global image features but also enhances the interactions between local regions of the image. Therefore, we adopts the MANIQA, which is the latest IQA approach, to evaluate image quality, in order to mitigate the negative effect of low-quality images on the model performance.

2.3 Cross-modal Matching

Cross-modal matching aims to align data from different modalities to measure similarity. Early works typically focus on coarse-grained alignment [6]. Some prior research [12] adopt attention mechanisms to achieve fne-grained local alignments. Recently, with the success of transformer-based vision-language pretrained models (VLP) [11, 19, 20, 9], such as CLIP [19], have shown strong performance in multiple cross-modal tasks. In the MNER-MI task, we observe that directly learning from images with weak textual correlations may introduce noisy interference, leading to suboptimal model performance. To address this, we employ CLIP to modeling cross-modal correlation, which implements dynamic sample weighting during training to mitigate interference from low-correlation images in model optimization.

3 Method

3.1 Problem Definition

MNER aims to utilize complementary visual information, such as images, to assist in recognizing and classifying named entities in text. In this paper, MNER is treated as a sequence labeling task. In the multimodal named entity recognition task, given a text sequence $X = x_1, x_2, ..., x_n$ and multiple associated images $I = I_1, I_2, ..., I_m$, where m represents the number of images (in our experiment, the maximum is 4 images). The goal is to recognize named entities from both text and images and assign them to predefined entity categories (e.g., Person, Location, Organization, and Miscellaneous). Specifically, the task is to assign a label sequence $Y = y_1, y_2, ..., y_n$ to each token in the text sequence, where each y_i belongs to an entity type set using the BIO (Beginning-Inside-Outside) labeling scheme.

3.2 Overview

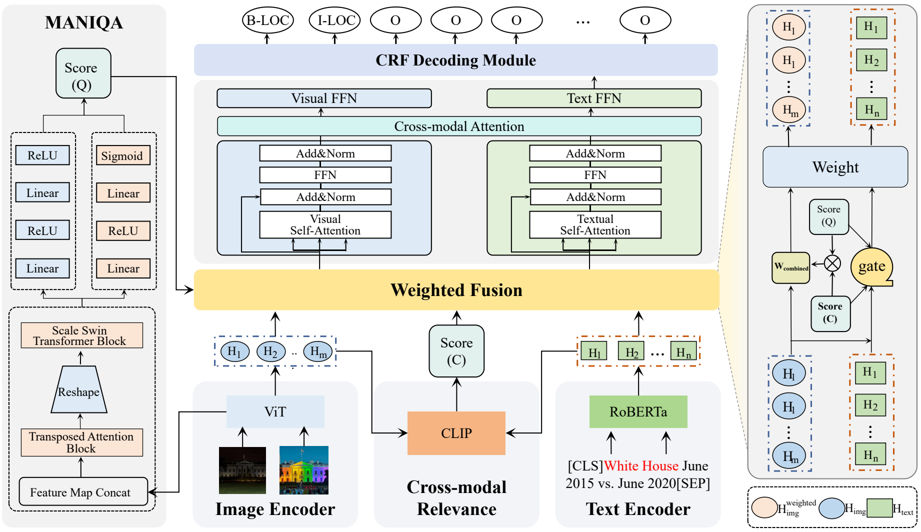

Fig. 2. Overall framework of IQCR-MI.

Our overall framework of IQCR-MI is shown in Figure 2, and the overall process is as follows.

We propose IQCR-MI, a novel multi-image integration framework for the multi-image MNER task. Current multi-image MNER approaches typically assume that all images contribute equally to the task. However, the quality of different images and their relevance to the text can significantly impact the task, and neglecting these factors may introduce noise, thus affecting the model's performance. To address this, we enhance the model's performance in multi-image tasks by incorporating image quality assessment and text-image relevance measurement.

The first component of the model is feature extraction, where we use RoBERTa as the text encoder to extract text features and Vision Transformer (ViT) to extract image features, providing the foundation for multimodal fusion. Next, we use the CLIP model to compute the semantic relevance between text and images, and assess image quality using the MANIQA model. Finally, we design a weighted fusion mechanism that combines image quality and text-image relevance scores to dynamically adjust the contribution of each image to entity recognition. The fused text and image features are then subjected to modality interaction, and entity label prediction is performed through a Conditional Random Field (CRF) layer.

3.3 Feature Extraction Module

Text Representation. To effectively extract semantic information from the text, we use RoBERTa [13] as the text encoder. RoBERTa is a Transformer-based pre-trained model with strong contextual understanding capabilities, enabling it to extract text representations that contain rich semantic information. It can also be replaced by other stronger text encoders. Given an input text sequence $X = x_1, x_2, ..., x_n$, we pass it through the RoBERTa model for encoding, obtaining the contextualized embedding representation of the text:

$$\mathbf{H}_{\text{text}} = \text{Encoder}(X) \tag{1}$$

Where, $\mathbf{H}_{\text{text}} \in \mathbb{R}^{n \times d_h}$ represents the contextualized token embeddings, where n is the length of the input text sequence, and d_h is the dimensionality of the embeddings.

Image Representation. For image feature extraction, we use the pre-trained ViT model [3]. ViT is a vision model based on the Transformer architecture, capable of capturing long-range dependencies within images and extracting high-quality visual features. Given an input image I, we use ViT to extract the global feature representation of the image:

$$\mathbf{H}_{\text{img}} = \text{ViT}(I) \tag{2}$$

For each image I,we extract the embedding of the [CLS] token as the global representation of the image. The resulting image feature representation is $\mathbf{H}_{\text{img}} \in \mathbb{R}^{m \times d_h}$. where m is the number of images, and d_h is the dimensionality of the image features.

3.4 Weighted Fusion Based on Image Quality and Cross-Modal Relevance

In multi-image scenarios, the quality differences of different images and their relevance to the text have a significant impact on entity recognition results. To address this, we design a weighted fusion method based on image quality and cross-modal relevance. Through the design of the following modules, we dynamically adjust the weight of each image, ensuring that high-quality and highly relevant images are better utilized in the model.

Image Quality Assessment. To effectively assess image quality, we adopt the MANIQA method. First, we preprocess each image to extract multiple local features $F_{i,j}$, where i denotes the image index and j represents the j-th local region of the image. These features are then concatenated. Next, the concatenated features pass through the Transposed Attention Block (TAB), where self-attention is applied across channels to capture global information. The image then proceeds to the Scale Swin Transformer Block (SSTB), which enhances local interactions between different regions of the image and improves the processing of fine-grained image features. Finally, the image features enter a dual-branch structure, which generates individual weights and quality scores for each patch of the image. These scores are then processed through the MANIQA model, yielding the final quality score Q_i for each image.The image quality score is calculated as follows:

$$Q_i = \text{MANIQA}(F_{i,1}, F_{i,2}, ..., F_{i,n}) \tag{3}$$

Where Q_i is the quality score of image I_i, and n is the number of local features extracted from the image. The value of Q_i reflects the visual quality of the image.

To address the quality differences across multiple images in a multi-image scenario, we generate a quality score matrix $\mathbf{Q}$ for all images in a batch, which represents the quality information of the images. The quality score matrix is given by: $\mathbf{Q} = \begin{bmatrix} Q_1 \; Q_2 \; \cdots \; Q_m \end{bmatrix}^T \in \mathbb{R}^{m \times n}$. Where m is the batch size, and n is the number of images.

Cross-modal Correlation Computation. To accurately compute the correlation between images and text, we adopt the CLIP model, which projects both text and images into a shared semantic space to measure their semantic similarity. CLIP is trained using contrastive learning, where it brings matched text-image pairs closer in the embedding space while pushing mismatched text-image pairs further apart.

For a given text x_i and image I_j, we first obtain their text embeddings c_i^{text} and image embeddings c_j^{img} using the CLIP model. The text and image embeddings represent the semantic information of the text and the image, respectively. Next, we compute the relevance score $S_{i,j}$ between the text and the image:

$$S_{i,j} = \frac{\langle c_i^{\text{text}}, c_j^{\text{img}} \rangle}{\|c_i^{\text{text}}\| \|c_j^{\text{img}}\|} \tag{4}$$

where $\langle c_i^{text}, c_j^{img} \rangle$ denotes the dot product between the text and image vectors,and denotes the dot product between the text and image vectors, and $\|c_i^{text}\|$ and $\|c_j^{img}\|$ represent their norms. The computed value $S_{i,j}$ falls within the range of [-1, 1], indicating the degree of correlation between the two.

Next, we compute the relevance score $S_{i,j}$ for each text-image pair and construct a relevance score matrix based on these scores: $\mathbf{C} = [C_1 \ C_2 \ \cdots \ C_n] \in \mathbb{R}^{m \times n}$. where m represents the batch size and n denotes the number of images. Each element $C_{i,j}$ in matrix C represents the semantic alignment between text x_i and image I_j. A higher relevance score indicates a stronger semantic relationship between the text and image, whereas a lower relevance score suggests a weaker connection.

Weighted fusion based on image quality and correlation. To effectively integrate image quality and text-image relevance, we designed a weighted fusion method based on image quality and relevance.

First, we normalize the image quality scores and text-image relevance scores to obtain the weight of each image:

$$W_{\text{qual}} = \frac{\exp(\alpha \cdot Q_i)}{\sum_{i=1}^{n} \exp(\alpha \cdot Q_i)} \tag{5}$$

$$W_{\text{corr}} = \frac{\exp(\beta \cdot C_j)}{\sum_{j=1}^{n} \exp(\beta \cdot C_j)} \tag{6}$$

where α and β are trainable scaling parameters that automatically adjust the importance of quality and relevance factors during the training process. Through softmax operation, we obtain the quality weight matrix W_{qual} and the relevance weight matrix W_{corr}, which respectively represent the influence of image quality and text-image relevance.

Then, through element-wise multiplication, we combine these two weight matrices to obtain the final weighted matrix:

$$\mathbf{W}_{\text{combined}} = \mathbf{W}_{\text{qual}} \odot \mathbf{W}_{\text{corr}} \tag{7}$$

where the matrix W_{combined} represents the overall influence of each image.

Next, we compute the gating value for each image to control its impact on the final recognition process. First, we perform a joint processing of the text features H_{text}, image features H_{img}, image quality scores Q, and text-image relevance scores C to obtain a gating value: $G \in [0, 1]$,where G denotes the sigmoid function. After processing of the multilayer perceptron (MLP), we obtain a gating value, which determines the influence of each image feature in the final fusion process.

Finally, we adjust the image features based on the gating value G to obtain the weighted image features:

$$H_{\text{img}}^{\text{weighted}} = G \cdot (W_{\text{combined}} \cdot H_{\text{img}}) + (1 - G) \cdot H_{\text{img}} \tag{8}$$

This design allows the model to retain the original image features while weighting the importance of features according to quality and relevance, forming a residual connection mechanism. When the image quality is low or the relevance to the text is weak, the gating value G will tend to be smaller, and the model will retain more original image features; when the image quality is high and highly relevant to the text, the gating value G will tend to be larger, and the model will use more weighted image features. In this way, we can dynamically adjust the contribution of each image in the final recognition based on its quality and relevance, ensuring that high-quality images with strong relevance to the text can occupy an important position in the model, thereby improving the accuracy of entity recognition.

3.5 Multi-modal Interaction Module.

To further enhance the deep interaction between text and images, we follow previous methods [29–32] and first adopt the self-attention Transformer layer to enhance the intra-modal interactions. Subsequently, we introduce the cross-modal Transformer layer to model the interaction between text and images.Specifically, the formulas are as follows:

$$\bar{\mathbf{H}}_{\text{text}} = \text{Self-ATT}(\mathbf{H}_{\text{text}}, \mathbf{H}_{\text{text}}, \mathbf{H}_{\text{text}}) \tag{9}$$

$$\bar{\mathbf{H}}_{\text{img}}^{\text{weighted}} = \text{Self-ATT}(\mathbf{H}_{\text{img}}^{\text{weighted}}, \mathbf{H}_{\text{img}}^{\text{weighted}}, \mathbf{H}_{\text{img}}^{\text{weighted}}) \tag{10}$$

$$\bar{\mathbf{H}}_{\text{fused}} = \text{Cross-ATT}(\bar{\mathbf{H}}_{\text{text}}, \bar{\mathbf{H}}_{\text{img}}^{\text{weighted}}, \bar{\mathbf{H}}_{\text{img}}^{\text{weighted}}) \tag{11}$$

where Self-ATT$(\cdot)$ represents single-modal multi-head self-attention mechanism [22], and Cross-ATT$(\cdot)$ represents cross-modal multi-head attention mechanism [21]. $\bar{\mathbf{H}}_{\text{text}}, \bar{\mathbf{H}}_{\text{img}}$, and $\bar{\mathbf{H}}_{\text{fused}}$ denote the final text representation, image representation, and word-level visual representation, respectively.

Finally, the weighted image and text features interact across multiple layers through self-attention and cross-attention mechanisms, ultimately obtaining rich multimodal representations.

3.6 CRF Decoding Module.

We employ CRF [10] as the decoder to perform conditional sequence labeling. CRF models the dependencies between adjacent labels and scores the entire label sequence. Specifically, we use linear-chain CRF to compute the conditional probability of the label sequence, formulated as follows:

$$p(y|\bar{\mathbf{H}}_{\text{fused}}) = \frac{\prod_{i=1}^{N} F_i(y_{i-1}, y_i, \bar{\mathbf{H}}_{\text{fused}})}{\sum_{y' \in Y} \prod_{i=1}^{N} F_i(y'_{i-1}, y'_i, \bar{\mathbf{H}}_{\text{fused}})} \tag{12}$$

where $F_i(y_{i-1}, y_i, \bar{\mathbf{H}}_{\text{fused}})$ and $F_i(y'_{i-1}, y'_i, \bar{\mathbf{H}}_{\text{fused}})$ are potential functions, which quantify the dependencies between labels.

Finally, we employ Maximum Conditional Likelihood Estimation (MLE) as the loss function of the model, defined as:

$$L(p(y|\bar{\mathbf{H}}_{\text{fused}})) = \sum_{i} \log p(y|\bar{\mathbf{H}}_{\text{fused}}) \tag{13}$$

This loss function maximizes the conditional probability of the ground-truth label sequence, allowing the model to more accurately predict the sequence labeling results, thereby enhancing NER performance. s

Table 1. Statistics of MNER-MI and MNER-MI-Plus.

Type	**MNER-MI**			**MNER-MI-Plus**		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
PER	4529	573	439	7472	1199	1060
LOC	1878	210	156	2609	383	334
ORG	1273	165	92	2947	540	487
MISC	2054	260	233	2755	410	390
Total	**9734**	**1208**	**920**	**15783**	**2532**	**2271**
Image	19188	2438	2395	22561	3161	3118
Tweet	6856	860	860	10229	1583	1583

4 Experiments

4.1 Experimental Setup

Datasets. This study uses MNER-MI and MNER-MI-Plus as evaluation datasets, where MNER-MI-Plus is an extension of MNER-MI that incorporates text-image pairs from Twitter 2017. A statistical analysis of the dataset is conducted to facilitate the advancement of multi-image MNER research. As shown in Table 1, the MNER-MI dataset consists of 6856 training tweets, 860 validation tweets, and 860 test tweets, while the MNER-MI-Plus dataset includes 10229 training tweets, 1583 validation tweets, and 1583 test tweets. The MNER-MI dataset contains only text-image pairs composed of multiple images, with 11,862 tweets and 24,021 images in tote. In contrast, the MNER-MI-Plus dataset, extended by the Twitter 2017 dataset, includes both text-image pairs composed of multiple images and text-image pairs composed of a single image.

Parameters Settings. Our experiments were conducted on an NVIDIA A800 GPU with 80GB of memory. We used PyTorch 2.5.1 and Transformers 4.46.3 as the deep learning framework, and CUDA 12.2 for GPU acceleration.For model configuration, we used Robert as the text encoder and ViT-base-patch16 as the image encoder. We used the AdamW optimizer with a learning rate of 2e-5 and set the batch size to 16. For evaluation metrics, we used Precision (P), Recall (R), and F1 score (F1)[29].

Baseline. In this study, our baseline methods include several classic pure text-based NER models. BiLSTM-CRF [8] combines Bidirectional Long Short-Term Memory (BiLSTM) and CRF to handle sequence labeling. CNN-BiLSTM-CRF [15] integrates Convolutional Neural Networks with BiLSTM to improve the model's performance. Additionally, although GPT-4, as a pretrained language model, is not specifically designed for NER, its powerful language understanding ability allows it to perform NER tasks effectively. BERT [2] uses the Transformer architecture for contextual embeddings, making it the standard model for modern NLP tasks.

In addition to these pure text models, we also employ several text+image models that combine image and text information for MNER tasks. Among these, MiniGPT-4 [35] is a multimodal large language model that enhances language understanding by inputting both text and images. GVATT-HBiLSTM-CRF [14] and AdaCAN-CNN-BiLSTM-CRF [33] utilize attention mechanisms to integrate image and text information for improved NER performance. UMT [29], MAF [26], UMGF [32], and HVPNeT [1] combine cross-modal learning with visual features to further enhance MNER performance.

For baseline comparison, this study adopts TPM-MI [7] as a benchmark model. This method is the first work on MNER with multiple images, proposing an approach that simulates multiple images as video frames and uses them as prompts to interact with the text for NER. Furthermore, to ensure a fair comparison, Huang et al. concatenated multiple images and validated single-image-based methods. Specifically, the study evaluated UMT-MI, UMGF-MI, and VisualPT-MoE-MI. Therefore, we also use these models as our comparative benchmarks to evaluate the effectiveness of the proposed method.

4.2 Experimental Results

As shown in Table 2, the IQCR-MI model outperforms the baseline methods on both datasets. Specifically, on the MNER-MI dataset, our method achieved a precision of 77.69%, recall of 79.44%, and an F1 score of 78.55%, surpassing the current state-of-the-art method, TPM-MI, with an improvement of 1.23% in the F1 score. Similarly, on the MNER-MI-Plus dataset, our method achieved a precision of 83.35%, recall of 86.25%, and an F1 score of 84.97%, surpassing TPM-MI by 1.55% in F1 score.

This performance improvement can be attributed to two key factors: (1) the integration of an image quality assessment mechanism, which effectively filters out low-quality images, thereby alleviating the noise caused by such images; (2) the adoption of a cross-modal relevance evaluation mechanism, which ensures that images highly relevant to the task provide more effective support for the NER task. Existing multi-image methods typically assume that all images contribute equally to the MNER task and overlook the impact of image quality on the task's outcome. In contrast, our IQCR-MI model dynamically adjusts the impact of each image on the MNER task by incorporating image quality assessment and cross-modal relevance mechanisms, further enhancing overall performance.

Table 2. performance comparison of our method with various baselines on the MNER-MI and MNER-MI-Plus datasets

Modality	**Models**	**MNER-MI**			**MNER-MI-Plus**		
		P	**R**	**F1**	**P**	**R**	**F1**
Text	BiLSTM-CRF	64.03	65.91	64.96	73.65	70.74	72.17
	CNN-BiLSTM-CRF	64.89	66.89	65.87	73.71	71.97	72.83
	GPT4	64.28	67.91	66.05	63.76	69.12	66.33
	HBiLSTM-CRF	64.51	68.55	66.47	72.19	74.34	73.25
	BERT	69.04	73.54	71.22	77.35	79.19	78.26
Text + Image	MiniGPT4	59.87	62.37	61.09	62.22	64.27	63.23
	GVATT-HBiLSTM-CRF	67.83	67.19	67.51	76.31	73.11	74.68
	AdaCAN-CNN-BiLSTM-CRF	67.89	68.24	68.06	75.67	73.85	74.75
	UMT	74.23	74.03	74.13	81.71	79.50	80.59
	MAF	74.91	73.60	74.25	80.17	81.29	80.73
	UMGF	73.74	75.30	74.51	82.31	79.65	80.96
	VisualPT-MoE	74.77	75.01	74.89	82.72	80.64	81.67
	HVPNeT	74.93	75.28	75.10	81.88	80.94	81.41
	UMT-MI	76.56	75.90	76.23	82.26	82.96	82.61
	UMGF-MI	75.88	77.14	76.50	82.55	82.25	82.40
	VisualPT-MoE-MI	76.87	76.38	76.62	82.61	82.79	82.70
	TPM-MI	77.45	77.19	77.32	83.66	83.18	83.42
	IQCR-MI (Ours)	**77.69**	**79.44**	**78.55**	83.35	**86.25**	**84.97**

5 Ablation Study

Table 3. Ablation study over two main components of proposed model

Model	**MNER-MI**			**MNER-MI-Plus**		
	P	**R**	**F1**	**P**	**R**	**F1**
w/o IQW	76.88	77.72	77.30	82.51	83.12	82.81
w/o CMW	77.14	78.04	77.59	83.26	83.42	83.34
IQCR-MI	77.69	79.44	78.55	83.35	86.25	84.97

To validate the effectiveness of each component in the IQCR-MI framework, we conducted comprehensive ablation experiments. Our approach includes several key components: Image Quality Weighting (IQW) and Cross-modal Weighting (CMW).

When the image quality weighting component (w/o IQW) was removed, we observed a 1.25% decrease in F1 score on MNER-MI and a 2.16% decrease on MNER-MI-Plus. This decline highlights the importance of the image quality mechanism in filtering out low-quality images, which can negatively impact the model's performance. As shown in Figure 1(a), when low-quality images are excluded, the performance improves, particularly when high-quality images are emphasized.

Removing the cross-modal weighting mechanism (w/o CMW) resulted in a 0.96% decrease in F1 score on MNER-MI and a 1.63% decrease on MNER-MI-Plus. This indicates that the semantic relevance between text and images plays a crucial role in performance. For instance, in Figure 1(b), the sentence "Ed Sheeran looks like the Stinky Cheese Man" loses its correct relationship with the image, as the cross-modal mechanism helps align the text and image content for better entity recognition.

These experiments demonstrate that the IQCR-MI framework, through its IQW and CMW significantly improves performance in the MNER-MI task. Each component contributes to the robustness of the model, addressing the challenges posed by multimodal inputs, and effectively fusing image and text information.

6 Further Analysis and Discussion

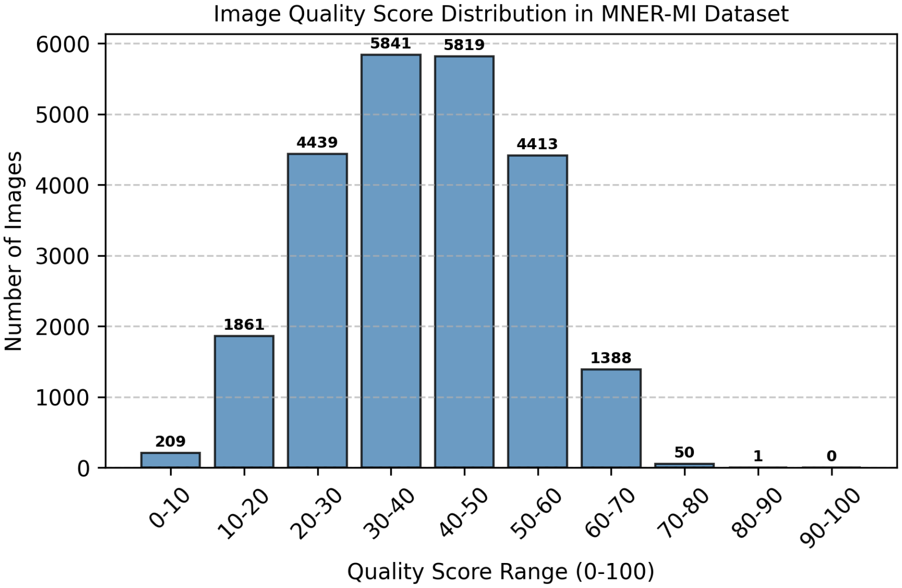

Fig. 3. Statistical Analysis of Image Quality Scores in the Multi-Image MNER Dataset.

In this section, we present a detailed statistical analysis of the image quality scores in the MNER-MI dataset.

First, we observe that images with quality scores above 50 account for only a small portion of the dataset, with 5,852 images, while images with scores below

50 make up the vast majority, totaling 18,170 images. This phenomenon clearly highlights the significant issue of image quality disparity in the current multi-image MNER tasks.

Secondly, there is only one image in the dataset with a quality score between 80 and 100. Despite the large number of images, high-quality images are extremely scarce. This finding suggests that the lack of high-quality images could limit the effectiveness of MNER tasks. Therefore, we propose that future research could focus on enhancing low-quality images to improve the overall image quality, which would likely boost the performance of MNER tasks.

7 Conclusion

In this paper, we propose a novel multi-image integration framework, named IQCR-MI, for the MNER-MI task. IQCR-MI dynamically adjusts the contribution of each image by incorporating image quality and cross-modal relevance, effectively guiding the fusion of features from multiple images. Our approach assigns higher weights to images with strong relevance while reducing the impact of low-quality or irrelevant ones, thereby alleviating performance bottlenecks caused by noisy or irrelevant image data. We experimented with IQCR-MI on two benchmark datasets, MNER-MI and MNER-MI-Plus, and demonstrated that our approach consistently outperforms existing multi-image methods, such as TPM-MI, achieving superior performance across precision, recall, and F1-score metrics.

References

1. Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., Huang, F., Si, L., Chen, H.: Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. arXiv preprint arXiv:2205.03521 (2022)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Gao, X., Gao, F., Tao, D., Li, X.: Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning. IEEE Transactions on neural networks and learning systems **24**(12) (2013)
5. Ghadiyaram, D., Bovik, A.C.: Perceptual quality prediction on authentically distorted images using a bag of features approach. Journal of vision **17**(1), 32–32 (2017)
6. Gur, S., Neverova, N., Stauffer, C., Lim, S.N., Kiela, D., Reiter, A.: Cross-modal retrieval augmentation for multi-modal classification. arXiv preprint arXiv:2104.08108 (2021)

7. Huang, S., Xu, B., Li, C., Ye, J., Lin, X.: Mner-mi: A multi-image dataset for multimodal named entity recognition in social media. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11452–11462 (2024)
8. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
9. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
11. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11336–11344 (2020)
12. Li, J., Liu, L., Niu, L., Zhang, L.: Memorize, associate and match: Embedding enhancement via fine-grained alignment for image-text retrieval. IEEE Transactions on Image Processing **30**, 9193–9207 (2021)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: 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). pp. 1990–1999 (2018)
15. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)
16. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
17. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity recognition for short social media posts. arXiv preprint arXiv:1802.07862 (2018)
18. Ok, H., Kil, T., Seo, S., Lee, J.: Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. arXiv preprint arXiv:2404.01914 (2024)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting. vol. 2019, p. 6558 (2019)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
23. Wang, X., Cai, J., Jiang, Y., Xie, P., Tu, K., Lu, W.: Named entity and relation extraction with multi-modal retrieval. arXiv preprint arXiv:2212.01612 (2022)

24. Wang, X., Tian, J., Gui, M., Li, Z., Ye, J., Yan, M., Xiao, Y.: Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In: International Conference on Database Systems for Advanced Applications. pp. 297–305. Springer (2022)
25. Wu, Z., Zheng, C., Cai, Y., Chen, J., Leung, H.f., Li, Q.: 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. pp. 1038–1046 (2020)
26. Xu, B., Huang, S., Sha, C., Wang, H.: 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. pp. 1215–1223 (2022)
27. Xu, B., Jiang, H., Wei, J., Jing, H., Du, M., Song, H., Wang, H., Xiao, Y.: Enhancing multimodal named entity recognition through adaptive mixup image augmentation. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 1802–1812 (2025)
28. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1191–1200 (2022)
29. YU, J., JIANG, J., YANG, L., XIA, R.: Improving multimodal named entity recognition via entity span detection with unified multimodal transformer.(2020). In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3342–3352 (2022)
30. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. Journal of machine learning research **3**(Feb), 1083–1106 (2003)
31. Zhang, D., Ju, X., Zhang, W., Li, J., Li, S., Zhu, Q., Zhou, G.: Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14338–14346 (2021)
32. Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G.: Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14347–14355 (2021)
33. Zhang, Q., Fu, J., Liu, X., Huang, X.: Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
34. Zheng, C., Wu, Z., Wang, T., Cai, Y., Li, Q.: Object-aware multimodal named entity recognition in social media posts with adversarial learning. IEEE Transactions on Multimedia **23**, 2520–2532 (2020)
35. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)