

MRE-MI: A Multi-image Dataset for Multimodal Relation Extraction in Social Media Posts

Shizhou Huang¹, Bo Xu², Changqun Li¹, Yang Yu¹, Xin Lin^{1,3,*}

¹School of Computer Science and Technology, East China Normal University

²School of Computer Science and Technology, Donghua University

³Shanghai Key Laboratory of Multidimensional Information Processing

huangshizhou@ica.stc.sh.cn, xubo@dhu.edu.cn,
{52215901009, 52205901014}@stu.ecnu.edu.cn, xlin@cs.ecnu.edu.cn

Abstract

Despite recent advances in Multimodal Relation Extraction (MRE), existing datasets and approaches primarily focus on single-image scenarios, overlooking the prevalent real-world cases where relationships are expressed through multiple images alongside text. To address this limitation, we present MRE-MI, a novel human-annotated dataset that includes both multi-image and single-image instances for relation extraction. Beyond dataset creation, we establish comprehensive baselines and propose a simple model named Global and Local Relevance-Modulated Attention Model (GLRA) to address the new challenges in multi-image scenarios. Our extensive experiments reveal that incorporating multiple images substantially improves relation extraction in multi-image scenarios. Furthermore, GLRA achieves state-of-the-art results on MRE-MI, demonstrating its effectiveness. The datasets and source code can be found at <https://github.com/JinFish/MRE-MI>.

1 Introduction

Multimodal Relation Extraction (MRE) has garnered significant attention for its ability to substantially enhance traditional text-based relation extraction by incorporating accompanying images as additional contextual information (Zheng et al., 2021a; Chen et al., 2022). This synergy between textual and visual information has demonstrated great potential in improving both the accuracy and scope of relation extraction across a variety of real-world scenarios.

According to Zhang et al. (2018), with the rise of social media, the trend of posts containing both text and multiple images has been steadily growing, with nearly half of all tweets now including multiple images. Despite this trend, current MRE datasets (Zheng et al., 2021b,a) and methods (Chen



Figure 1: Two examples of multimodal relation extraction with multiple images.

et al., 2022; Zheng et al., 2023; Liu et al., 2024) remain focused on text paired with a single image. For example, in the annotation process of MNRE (Zheng et al., 2021b), only one image is randomly selected in instances containing multiple images, thereby ignoring the valuable insights provided by the remaining images.

In addition, posts with multiple images usually require either **all images** or **key images** to accurately understand their meaning, making it challenging for current MRE methods that consider only a randomly selected single image to effectively handle multi-image scenarios. This characteristic is also present in multi-image multimodal named entity recognition (Huang et al., 2024). As shown in the first example in Figure 1, it is difficult to determine the relation between *Stephen curry* and *LeBron James* when only text and either of the two images are considered. With the help of **all the images**, we can judge the relation between the two as *Competitor* by their different jerseys. As illustrated in the second example in Figure 1, it is challenging to determine the relation between *Hailey Baldwin* and *Justin Bieber* when there is only text and the first two images. With the help of the third image (**key image**), we can identify their relation as *Couple* by the action of kissing in the

*Corresponding Author.

image.

To explore MRE in scenarios with multiple images, we introduce a novel human-annotated dataset, named **MRE-MI** (Multimodal Relation Extraction with Multiple Images) in this paper. The MRE-MI dataset is collected from Twitter and each instance contains at least 1 image and at most 4 images (the maximum number of images allowed per post on Twitter). This dataset contains more data (22,504 instances) compared to the existing MRE dataset, as well as more labels (26 labels) to expand the range of scenarios covered.

While utilizing multiple images offers rich contextual information, it also introduces new challenges, such as representing and interacting with text and multiple images, as well as determining the importance and contribution of each image. To address these challenges, we propose a simple yet powerful model called the Global and Local Relevance-Modulated Attention Model (GLRA). This model establishes relationships between text and multiple images by capturing both global and local representations of text and images. It evaluates the relevance between text and images to determine the importance of each image, incorporating this relevance into the attention mechanism to adjust the fusion of text and images.

Moreover, we establish a diverse and representative set of baselines, including text-based RE methods, MRE methods, and Large Language Models (LLMs). We compare the performance of these methods on both single-image and multi-image instances in MRE-MI. While current single-image MRE methods perform well in single-image scenarios, they struggle in multi-image contexts. This highlights the limitations of existing single-image MRE methods in handling multi-image scenarios, their lack of direct applicability to multi-image scenes, and the complexity and challenges in our proposed dataset.

Our main contributions are summarized as follows:

- We present MRE-MI, a novel and challenging human-annotated dataset designed to bridge the research gap in MRE and to broaden its applicability to real-world applications.
- We establish a comprehensive set of strong and representative baselines on MRE-MI. Experimental results demonstrate that incorporating multiple images substantially improves model performance in multi-image contexts

compared to using only a single image, highlighting the advantages of multiple images in better understanding multimodal content.

- To address the challenges in multiple images, we propose GLRA, and experimental results demonstrate that GLRA achieves state-of-the-art results performance on MRE-MI, demonstrating its effectiveness in handling multi-image scenarios.

2 Related Work

2.1 Multimodal Relation Extraction

In this paper, we focus on MRE for images and text. Most methods (Zheng et al., 2021a; Chen et al., 2022; Xu et al., 2023; Liu et al., 2024) directly use a language model (e.g., BERT) to obtain the text representation and use various visual models to obtain the image representation. For example, Zheng et al. (2021b) uses a scene graph model to obtain the relations between objects in the image as the image representation. Chen et al. (2022); Wei et al. (2024); Liu et al. (2024) use a visual grounding model to obtain image regions related to the text and then input these regions and the original image into an image encoder to get the image representation.

For the interaction of text and images, current methods are mainly based on the attention mechanism. Early work (Zheng et al., 2021b,a) directly interacts with text and images through an attention mechanism. Subsequently, Chen et al. (2022); Xu et al. (2023) project image representation as the prompts, allowing it to interact with the text representation through the attention layer of the text encoder. In addition, Chen et al. (2022); Liu et al. (2024) implements the interaction of different levels of text and image representations based on the attention mechanism.

Furthermore, some methods focus on the reduction of the effect of image noise, Xu et al. (2022a) applies reinforcement learning to train a data discriminator to determine the relevance between text and images. Chen et al. (2022) proposes to project images as prompts to serve as prefixes for the text and Xu et al. (2023) proposes to add a pseudo-image representation to mitigate the impact of image noise.

Additionally, Sun et al. (2024) proposes a unified information extraction framework that models MRE as a generation problem.

2.2 Datasets for Multimodal Relation Extraction

To the best of our knowledge, Zheng et al. (2021b) is the first to propose the MRE dataset, which contains data from Twitter. In this dataset, when a tweet includes more than one image, one is selected randomly. Zheng et al. (2021a) then improves the dataset by merging some ambiguous categories to create a dataset called MNRE, which is also the most widely used resource in the MRE task.

Given that the existing dataset primarily combines text with a single image, there remains a notable research gap in handling scenarios involving multiple images. In addition, recent work on multimodal named entity recognition in multi-image scenarios (Huang et al., 2024) has gained attention and demonstrated the advantages of considering multiple images in such scenarios. Therefore, to address this problem and to fully utilize multiple images in social media, we propose the multi-image MRE dataset MRE-MI.

3 Datasets

3.1 Dataset Collection

Since the current single-image MRE dataset MNRE (Zheng et al., 2021a) does not provide a detailed annotation guidelines document, we have opted not to include MNRE data in our dataset. This decision prevents issues related to inconsistencies between the annotation rules of the new data and those of MNRE.

We follow Lu et al. (2018); Zheng et al. (2021b) to collect tweets from Twitter¹. Unlike MNRE, we do not restrict our data collection to a fixed number of months in a single year. Instead, we gather tweets for each month from 2020 to 2023, resulting in a more diverse and unbiased dataset, which also presents greater challenges. We then filter out non-English tweets, repeated tweets, tweets with a text length of less than 3 words, and tweets without images. Ultimately, we get the 10K+ tweets for annotation, where each tweet contains up to 4 images (the maximum number of images on Twitter).

3.2 Human Annotation

We employ five well-educated annotators to annotate the tweets. Each annotator can view the text and all the images in the tweet during the annotation process and uses them to label the categories

between entities. We aggregate the annotations using majority voting.

Additionally, if any annotator identifies a tweet as revealing personal information or containing sensitive or harmful content, that tweet is discarded. We adopt Fleiss Kappa (Fleiss, 1971) to measure the annotation agreement, and the Fleiss score between the three annotators is $\mathcal{K} = 0.84$, indicating a substantial annotation agreement.

Finally, we obtain 9,652 tweets and label 22,504 entity-pair relations from these tweets. We divide our dataset into the train (60%), validation (20%), and test sets (20%). There are 13,504/4,500/4,500 entity pairs in the train/development/test set, respectively.

Statistics	MNRE	MRE-MI (Our Dataset)
# Word	258k	363k
# Instance	15,485	22,504
# Relation	23	26
Scenarios	SI	SI & MI

Table 1: A comparison with MRE dataset. SI and MI represent that this dataset is used for single-image scenarios and multi-image scenarios, respectively.

3.3 Dataset Analysis

As shown in Table 1, we compare the statistics of our dataset with those of MNRE, highlighting that our dataset incorporates multi-image scenarios, enhancing its relevance to real-world applications. Additionally, MRE-MI features a greater number of sentences, instances, and relations, resulting in a richer and more detailed dataset. This enhancement broadens the range of scenarios and increases the dataset’s capacity to capture diverse relations.

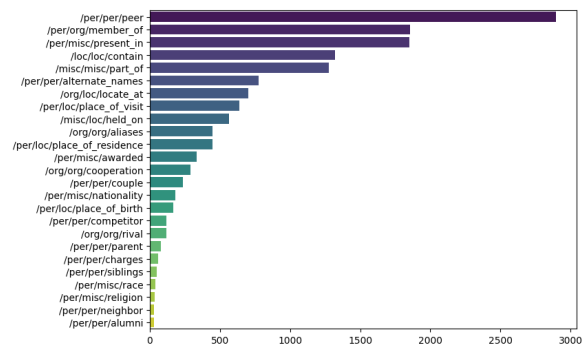


Figure 2: The distribution of relation categories in our MRE-MI dataset.

We also present the distribution of the 25 relation categories in our MNRE-MI dataset in Figure 2.

¹<https://archive.org/details/twitterstream>

Entity pairs that do not belong to these relation categories are classified as *None*. Following (Zheng et al., 2021a), we prepend the entity types of the two entities involved in the relations to the labels. This approach facilitates the use of entity categories to aid in relation extraction in future work. Note that the entity types are included in the labels without disclosing the actual entity information.

4 Method

In this section, we first formulate our problem, and then describe the main components in our proposed model GLRA: (1) Text Representations, (2) Image Representations, and (3) Relevance-Modulated Cross-Modal Attention.

4.1 Problem Formulation

Given a text $S = (s_1, s_2, \dots, s_n)$, marked entities e_1 and e_2 within text, and associated images $I = (i_1, i_2, \dots, i_m)$ as input, where n is the number of tokens in the text, and m is the number of images. The objective is to classify the corresponding relation Y between the entities e_1 and e_2 .

4.2 Overall Framework

Our overall framework of GLRA is shown in Figure 3, and the overall process is as follows.

For text representations, we first preprocess the text by using special tokens to mark the positions of the two entities. The processed text is then fed into BERT (Devlin et al., 2019) to obtain the representation for each token. The representation of the [CLS] token serves as the global text representation, while the representations of the special tokens serve as the local text representation.

For image representations, we first use a visual grounding tool (Yang et al., 2019) to identify the image regions related to the text in each image. We then input multiple original images and image regions into ViT (Dosovitskiy et al., 2020) to separately obtain the global image representation and the local image representation.

Next, we feed the global text representation and the global image representation into our proposed relevance-modulated attention module to obtain the global text-aware image representation, which uses the relevance between text and images to adjust the attention scores between them.

Similarly, we input the local text representation and the local image representation into the same module to obtain the local text-aware image representation. Finally, we concatenate the global and

local text representations with the text-aware image representations to predict the relations of entities.

4.3 Text Representations

In this paper, we use BERT (Devlin et al., 2019) as the text encoder. Firstly, we follow Devlin et al. (2019) and add a [CLS] token and a [SEP] token at the beginning and end of the text input S to represent the start and end of the text. Then, we follow previous work (Wei et al., 2024; Liu et al., 2024) and add special tokens $\langle s \rangle$, $\langle /s \rangle$, $\langle o \rangle$, $\langle /o \rangle$ to indicate the start and end of the two entities. The final processed text input T is as follows:

$$T = ([CLS], t_1, \dots, \langle s \rangle, t_{e_s}, \langle /s \rangle, \dots, \langle o \rangle, t_{e_o}, \langle /o \rangle, \dots, t_n, [SEP]) \quad (1)$$

where t_{e_s} and t_{e_o} are tokenized sub-sequences of two entities, respectively.

Finally, we feed the T to BERT to obtain the representation of token sequence $\mathbf{H} \in \mathbb{R}^{n_t \times d_t}$, where n_t is the sequence length and d_t is the dimension of the text representations. We use the representation of [CLS] as the global text representation $\mathbf{H}_{[CLS]} \in \mathbb{R}^{1 \times d_t}$, and use the entity representation of $\langle s \rangle$ and $\langle o \rangle$ as the local text representations: $[\mathbf{H}_{\langle s \rangle}; \mathbf{H}_{\langle o \rangle}] \in \mathbb{R}^{2 \times d_t}$.

4.4 Image Representations

In this paper, we use ViT (Dosovitskiy et al., 2020) as the image encoder. Firstly, we follow Chen et al. (2022); Wei et al. (2024); Liu et al. (2024) and adopt a visual grounding tool (Yang et al., 2019) to extract the top k salient image regions in each of m input images. Then, we input them into ViT, obtaining global image representations $\mathbf{V} = (v_1, \dots, v_m) \in \mathbb{R}^{m \times d_v}$ and local image representations $\mathbf{O} = (o_1, \dots, o_k, o_{k+1}, \dots, o_{m \times k}) \in \mathbb{R}^{(m \times k) \times d_v}$, respectively, where d_v is the dimension of the text representations.

4.5 Relevance-Modulated Cross-Modal Attention

The native cross-modal attention mechanism (Yu et al., 2020; Xu et al., 2022b) essentially distributes text weights across all images, meaning that even if all images contain significant noise, the weights are still distributed among them, potentially introducing more noise in multi-image scenarios. Moreover, in the case of a single image, all weights are assigned to that image, which retains all its information and can render the attention mechanism ineffective.

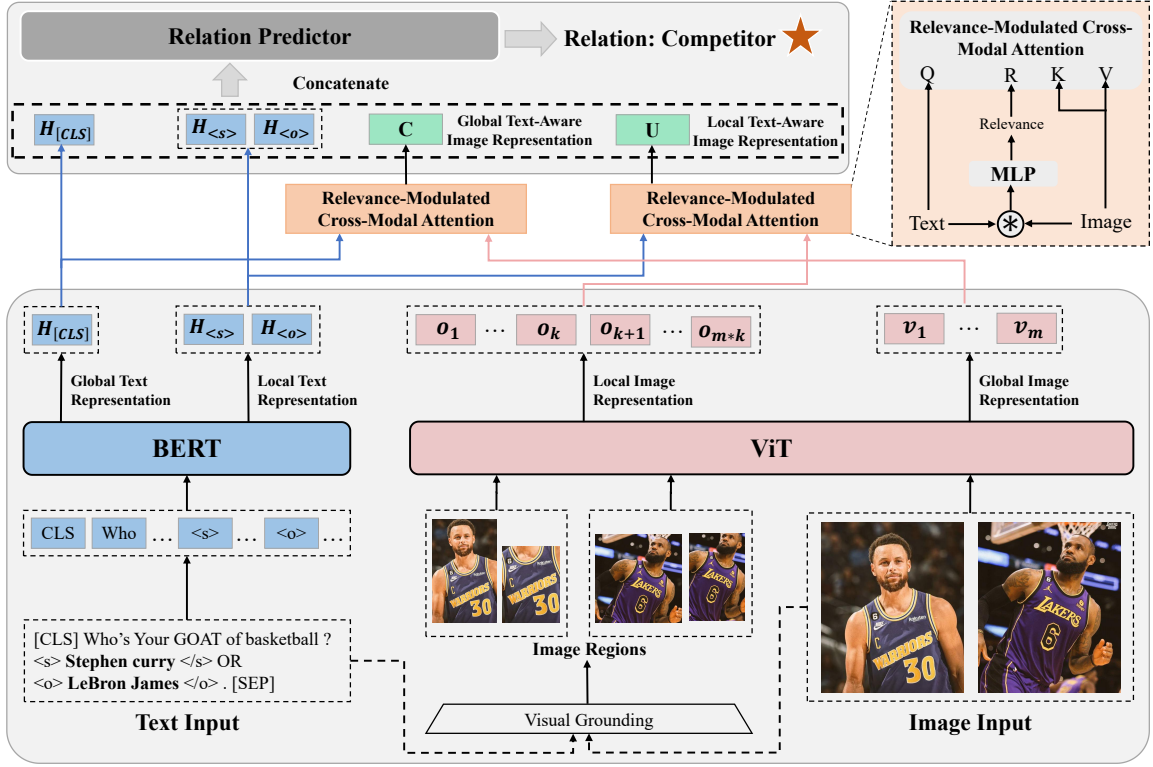


Figure 3: Overall framework of GLRA.

To address this issue, we propose a method called **Relevance-Modulated Cross-Modal Attention**. This method first calculates the relevance between text and images independently and then uses this relevance to adjust the attention scores. We illustrate the process using global text and image representations as examples. Specifically, we first project these representations into the same space:

$$\hat{\mathbf{H}}_{[\text{CLS}]} = \mathbf{H}_{[\text{CLS}]} \mathbf{W}_1 \quad (2)$$

$$\hat{\mathbf{V}} = \mathbf{V} \mathbf{W}_2 \quad (3)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_t \times d_t}$, $\mathbf{W}_2 \in \mathbb{R}^{d_v \times d_t}$ are the weight matrices.

Then, following (Xu et al., 2022a), we capture the relevance features between text and images through element-wise multiplication, and input these features into an MLP to obtain the final relevance, which offers a more flexible integration of relevance features compared to cosine similarity:

$$\mathbf{R}_{\text{global}} = \text{sigmoid}(\text{relu}((\hat{\mathbf{H}}_{[\text{CLS}]} \odot \hat{\mathbf{V}}) \mathbf{W}_3) \mathbf{W}_4) \quad (4)$$

where \odot denotes the element-wise multiplication, $(\hat{\mathbf{H}}_{[\text{CLS}]} \odot \hat{\mathbf{V}}) \in \mathbb{R}^{1 \times m \times d_t}$ denotes the relevance feature between a text representation and m image representations, sigmoid and relu are the activation functions, $\mathbf{W}_3 \in \mathbb{R}^{d_t \times d_t}$ and $\mathbf{W}_4 \in \mathbb{R}^{d_t \times d_1}$

are weight matrices, $\mathbf{R}_{\text{global}} \in \mathbb{R}^{1 \times m}$ representing the relevance score between the global text representation and the m global image representations, with the relevance ranging between 0 and 1.

Next, we use the $\mathbf{R}_{\text{global}}$ to modulate the attention scores in the original attention mechanism and obtain the global text-aware image representation:

$$\alpha = \text{softmax}\left(\frac{[\mathbf{H}_{[\text{CLS}]} \mathbf{W}_q][\mathbf{V} \mathbf{W}_k]^T}{\sqrt{d_t}}\right) \quad (5)$$

$$\mathbf{C} = (\mathbf{R}_{\text{global}} \odot \alpha) [\mathbf{V} \mathbf{W}_v] \quad (6)$$

where $\mathbf{W}_q \in \mathbb{R}^{d_t \times d_t}$ and $\{\mathbf{W}_k, \mathbf{W}_v\} \in \mathbb{R}^{d_v \times d_t}$ are weight matrices, $\alpha \in \mathbb{R}^{1 \times m}$ is the attention scores between a global text representation and m global image representations, $\mathbf{C} \in \mathbb{R}^{1 \times d_t}$ is the global text-aware image representation.

We follow the above approach to obtain the local text-aware image representation $\mathbf{U} \in \mathbb{R}^{2 \times d_t}$.

4.6 Predictor

We concatenate the text and image representations together and input them into a linear layer with a softmax activation function to predict the probability assigned by the model to the true label:

$$p(y|S, I) = \text{softmax}([\mathbf{H}_{[\text{CLS}]}; \mathbf{H}_{<s>}; \mathbf{H}_{<o>}; \mathbf{C}; \mathbf{U}] \mathbf{W}) \quad (7)$$

where $\mathbf{W} \in \mathbb{R}^{(6*dt) \times |Y|}$ is the weight matrix, $|Y|$ is the number of categories.

The model parameters are optimized by minimizing the cross-entropy loss as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{j=1}^N \log(p(y^j | S^j, I^j)) \quad (8)$$

where N is the batch size.

5 Experiments

In this section, we conduct various experiments to comprehensively evaluate the performance of baselines and our methods on MRE-MI. Following many works (Chen et al., 2022; Xu et al., 2023; Liu et al., 2024), we use the accuracy (**Acc.**) and F1 score (**F1**) as evaluation metrics.

5.1 Baselines

Text-based models: For text modality, we explore several models commonly used in relation extraction tasks: (1) PCNN (Zeng et al., 2015) is a distantly supervised model that leverages external knowledge to label sentences with the same entities contained automatically; (2) BERT, a powerful transformer-based text encoder; (3) MTB (Soares et al., 2019) is a RE-oriented model based on BERT that introduces special tokens to denote the entity positions.

Multimodal relation extraction models: For our multimodal experiments involving both text and image modalities, we use the current representative MRE models as baselines. Specifically, (1) UMT (Yu et al., 2020) proposes a multimodal interaction module for establishing bi-directional relationships between text and images; (2) UMGF (Zhang et al., 2021) proposes an approach based on a graph model to establish the relationship between text and images; (3) VisualBERT (Li et al., 2019) is a single-stream multimodal pre-trained model; (4) CLIP (Radford et al., 2021) is a dual-stream multimodal pre-trained model that learns visual and textual representations jointly in a shared embedding space; (5) MEGA (Zheng et al., 2021a) is a model with efficient graph alignment; (6) HVPNeT (Chen et al., 2022) and (7) VisualPT-MoE (Xu et al., 2023) all project image representations as prompts to achieve interaction with each layer of the text encoder; (8) UMIE (Sun et al., 2024) proposes a generation model as an extractor to classify the relation between entities; (9) CGI-MRE (Wei et al., 2024) is a

model that applies genetic concepts to MRE; (10) HVFormer (Liu et al., 2024) integrates different levels of text representation and image representation. All of the above methods are single-image MRE methods, so for convenience, we select the first image as their image input.

Additionally, both HVPNeT and HVFormer can accept multiple images as input, leading to the development of (11) HVPNeT-MI and (12) HVFormer-MI, respectively. (13) TPM-MI (Huang et al., 2024) is a model designed for multi-image multimodal named entity recognition, treating multiple images as frames of a video; (14) GLRA is the approach we propose in this paper. All of the above four methods are multi-image MRE methods.

Large language models: Considering the advancements in large language models that can perform various tasks in a zero-shot manner. We use the current advanced large language model, Qwen2-VL (Qwen2-VL-72B-Instruct) (Wang et al., 2024), LLaVA-OV (llava-onevision-qwen2-72b-ov-hf) (Liu et al., 2024) and GPT-4o, as a baseline, which can be applied in single-image and multi-image scenarios.

5.2 Experimental Settings

All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs (24GB) with PyTorch 2.1.2, and the parameters settings of our model and baselines are as follows:

- We use BERT-base and ViT-large-patch16-384 as the text encoder and image encoder for all methods, respectively.
- We use the AdamW as the optimizer and use the grid search in the development set to find the learning rate within $[1e^{-5}, 7e^{-5}]$, the batch size within $[8, 32]$. Our approach sets the final learning rate to $4e^{-5}$ and the batch size to 16.
- The training process employs mini-batch back-propagation and the maximum number of training epochs is set to 20. We select the model that performs best on the development set and evaluate it on the test set.

5.3 Main Results

As shown in Table 2, we first compare all text-based methods. We find that MTB significantly outperforms BERT, highlighting the effectiveness of using markers to indicate entity positions. Next, we

Modality	Model	MRE-MI	
		Acc.	F1
Text	PCNN	49.93	53.12
	BERT	52.58	55.21
	MTB	54.38	57.07
Text + One Image	UMT	55.23	57.82
	UMGF	56.43	58.17
	VisualBERT	52.31	54.69
	CLIP	54.62	55.76
	MEGA	56.76	58.73
	HVPNeT	57.50	59.98
	VisualPT-MoE	57.26	59.53
	UMIE	53.32	55.27
	CGI-MRE	57.65	60.34
	HVFormer	58.44	60.82
	Qwen2-VL	45.32	47.68
	LLaVA-OV	47.79	49.36
	GPT-4o	48.73	50.75
Text + All Image	HVPNeT-MI	58.67	62.42
	HVFormer-MI	59.48	62.67
	TPM-MI	58.01	61.34
	Qwen2-VL	49.78	51.22
	LLaVA-OV	50.94	52.63
	GPT-4o	52.06	55.06
	GLRA	61.60[†]	65.54[†]

Table 2: Performance comparison on MRE-MI. The marker [†] refers to significant test p-value < 0.05 when compared with HVFormer-MI.

compare multimodal methods of using the single image with the text-based methods. We find that almost all multimodal models significantly outperform their corresponding text-based models, such as UMT vs. BERT, and HVPNeT vs. MTB. This indicates that image information in social media posts contributes to relation extraction.

However, we find that multimodal LLMs fail to achieve satisfactory results compared to fine-tuning methods, and this phenomenon is also found in text-based relation extraction dataset (Li et al., 2023). This highlights the significant challenges information extraction tasks pose for LLMs. (Pang et al., 2023) attributes this to the complexity of the information extraction task, where LLMs often struggle to align their understanding of labels with the actual meanings of labels in the dataset. For example, in the MNRE dataset, the relation between the entity Ruby Rose and Batwoman in the text “The CW has released the first look at Ruby Rose as Batwoman” is classified as /per/per/alternate_names. However, LLMs often do not consider the relation between an actor and the character they portray as alternate_names and are more likely to classify it as None.

Finally, we compare multimodal methods of using multiple images. Our findings show that these methods consistently outperform their single-image counterparts, indicating that leveraging multiple images in multi-image scenarios enhances the understanding of multimodal content and facilitates relation extraction. Notably, Qwen2-VL, LLaVA-OV and GPT-4o exhibit the most significant improvement, which we attribute to the robust capabilities of large language models in processing multiple images. This allows them to effectively harness image information for complex reasoning tasks. However, they still exhibit a significant gap compared to fine-tuning methods. Furthermore, we find that our proposed methods GLRA significantly outperform the other methods, which demonstrates the effectiveness of our method.

Model	MRE-MI	
	Acc.	F1
GLRA	62.87	65.76
w/o Global Representations	61.01	63.56
w/o Local Representations	58.33	60.21
w/o Relevance Modulation	60.14	62.27

Table 3: Ablation study of our GLRA. w/o Relevance Modulation means that only the relevance-modulated part is removed and the attention is still retained.

5.4 Ablation Study

To investigate the effectiveness of the global representations (global text and image representation), local representations (local text and image representation) and relevance-modulated cross-modal attention, we perform comparisons between the full model GLRA and its ablation methods.

As shown in Table 3, GLRA benefits from all the aforementioned components. Specifically, we find that removing local representations impairs the performance of model more than removing global representations, indicating that MRE relies more on the representations of entities in the text and images. Additionally, we observe a significant performance drop when the relevance modulation part is removed, demonstrating that prior computation of the relevance between text and image can effectively reduce the impact of image noise.

5.5 Performance Under Different Dataset Settings

In Figure 4, we show the performance of the models on the MRE-MI dataset under different settings.

Our analysis reveals that while models perform well on single-image instances, their performance significantly drops on multi-image instances, with a maximum degradation exceeding **30 points**. This indicates that current single-image MRE methods are not directly applicable to multi-image scenarios. Simply adapting existing methods for multi-image contexts does not yield satisfactory results, underscoring the unique challenges our dataset presents compared to single-image MRE datasets.

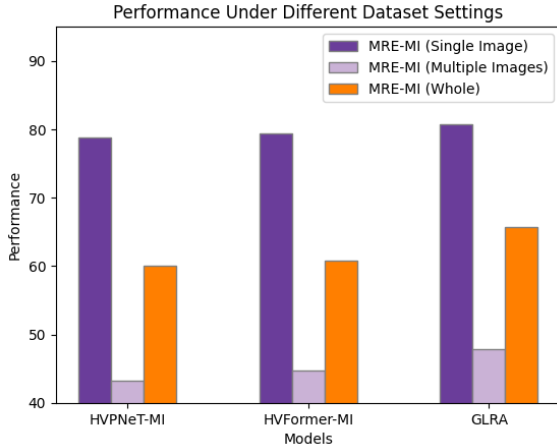


Figure 4: Performance (F1) comparison on different datasets. Single Image denotes instances of a single image on the dataset, Multiple Images denotes instances of multiple images on the dataset, and Whole denotes all instances on the whole dataset.

5.6 Case Study

To more visually show the effectiveness of using multiple images in a multi-image scenario, we conduct a case study and compare the different methods as shown in Table 4.

Specifically, in the first example, with the help of multiple images, HVFormer-MI and GLRA can determine the relation between Sofia Carson and Nicholas Galizine by considering the people in the images. However, HVFormer, which only uses the first image, does not obtain enough information and incorrectly classifies their relation as peer. This illustrates that using multiple images in a multi-image scenario can be more effective in determining relations between entities.

In the second example, the first image is not helpful in determining the relations between the entities. Affected by the noise in the first image, HVFormer and HVFormer-MI incorrectly predict entity relations that could be judged by text alone. In contrast, GLRA reduces the noise introduced

by images and predicts the correct relation, demonstrating the effectiveness of our approach.

Furthermore, in the third example, the text in the images indicates that John Wick and Extraction are movie titles, leading to the inference that Chris Hemsworth is involved in Extraction. However, current MRE methods lack the capability for complex multimodal reasoning, and all three models incorrectly predict the relation between entities. This indicates that the current methods have obvious limitations in requiring reasoning about text and image information.

6 Conclusion

In this paper, in order to address the research gaps in MRE as well as to expand the scope of MRE for real-world applications, we propose a multi-image MRE dataset MRE-MI. In addition, we establish a comprehensive set of representative baseline methods and propose a novel model for the challenges of multi-image relation extraction. We have conducted extensive experiments to demonstrate that multiple images can provide more information to better help MRE compared to a single image, and the effectiveness of our method.

7 Limitations

Although our work has made progress and we believe that the dataset we have proposed could be a valuable resource for the research community. However, since this is a resource paper, our main goal is to release the dataset, with only a preliminary exploration of methods for this task, there are some limitations:

Firstly, our method requires identifying the image regions where objects are located in each image to obtain more fine-grained visual representations for alignment with entities in the text, resulting in the generation of many images, increasing the model’s runtime.

Secondly, similar to previous work, our method only establishes correspondences between entities in the text and images, lacking multimodal reasoning capabilities, leading to poor performance in scenarios that require inference.

Lastly, we have not proposed solutions for some challenges in multi-image scenarios. For example, our method does not model the relationships between images, whereas considering the relationships between images could better help in understanding the overall meaning of the post.


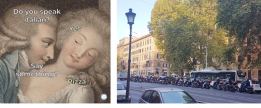

Sentence and Images	Label	HVFormer	HVFormer-MI	GLRA
The chemistry between Sofia Carson and Nicholas Galitzine in Purple Hearts . 	couple	peer ×	couple ✓	couple ✓
When I first came to Rome, Italy . 	contain	couple ×	couple ×	contain ✓
In 2023, John wick Or Chris Hemsworth's Extraction 	present_in	awarded ×	peer ×	peer ×

Table 4: A case study, where bold text is the entity, HVFormer uses the first image, and others use multiple images.

In the future, we will focus on developing new models with better performance on MNER-MI to address the above limitations. Furthermore, we hope that the release of this dataset will stimulate more research in the field, leading to the development of innovative models.

8 Ethical Considerations

As mentioned in Section 3.2, we discard tweets containing personal information, as well as any sensitive or harmful content. We used an AI assistant (ChatGPT) to help us refine the rules for determining whether personal privacy has been leaked and whether harmful or sensitive information is included. As a result, We do not share personal information and do not release sensitive content that can be harmful to any individual or community.

Prior to annotation, we inform the annotators of the tasks they will be performing, the data to be annotated, the approximate amount of data each individual will need to annotate, and the cycle time, as well as the risk of exposure to harmful, sensitive text or images. Moreover, all annotators can quit their jobs at any time.

To the best of our knowledge, we are not aware of any other possible ethical consequences of the proposed dataset.

9 Acknowledgments

This work is supported by National Science and Technology Major Project

(2021ZD0111000/2021ZD0111004), the Science and Technology Commission of Shanghai Municipality Grant (No. 21511100101, 22511105901, 22DZ2229004), the Fundamental Research Funds for the Central Universities 2232023D-19, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education. Xin Lin is the corresponding author. Xin Lin is also a member of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education.

References

- Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,

- Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Shizhou Huang, Bo Xu, Changqun Li, Jiabo Ye, and Xin Lin. 2024. 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)*, pages 11452–11462.
- Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt’s information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Xiyang Liu, Chunming Hu, Richong Zhang, Kai Sun, Samuel Mensah, and Yongyi Mao. 2024. Multimodal relation extraction via a mixture of hierarchical visual context learners. In *Proceedings of the ACM on Web Conference 2024*, pages 4283–4294.
- Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.
- Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15372–15389.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.
- Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19062–19070.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Pengfei Wei, Zhaokang Huang, Hongjun Ouyang, Qintai Hu, Bi Zeng, and Guang Feng. 2024. Cgi-mre: A comprehensive genetic-inspired model for multimodal relation extraction. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 524–532.
- Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Chaofeng Sha, and Yanghua Xiao. 2022a. Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1855–1864.
- Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Yanghua Xiao, and Xin Lin. 2023. A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction. In *International Conference on Database Systems for Advanced Applications*, pages 544–554. Springer.
- Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022b. 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.
- Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.
- Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.

- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Changmeng Zheng, Junhao Feng, Yi Cai, Xiaoyong Wei, and Qing Li. 2023. 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.
- Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306.
- Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.