Text-Guided Dual Interaction for Multimodal Relation Extraction in Social Media

Yachuan Zhang[1] and Yi Guo[1,2,3] (✉)

[1] East China University of Science and Technology, Shanghai 200237, China
Y10220122@mail.ecust.edu.cn, guoyi@ecust.edu.cn
[2] Shanghai Engineering Research Center of Big Data and Internet Audience, Shanghai, China
[3] Business Intelligence and Visualization Research Center, National Engineering Laboratory for Big Data Distribution and Exchange Technologies, Shanghai, China

Abstract. Multimodal relation extraction is essential for information extraction and knowledge graph construction. In social media, in some situations, text and images often lack relevance or have weak connections, which can mislead models. While many current approaches focus on modality alignment and fusion, they overlook the role of domain-specific modality in mitigating information bias. Moreover, significant gaps between modalities make it challenging to establish deep associative relationships. To tackle these challenges, we propose the Text-Guided Dual Interaction (TGDI) model, which incorporates a Modal Dual-Interaction mechanism. Specifically, the Cross-Modal Interaction module performs global level fusion to achieve initial alignment, while the Text-Oriented Interaction module refines this integration by preserving essential visual information under textual guidance. Additionally, the Text Modulated Matching Gate regulates visual contributions and evaluates image-text similarity to minimize visual noise. Finally, the fusion function adapts to various text-image scenarios, ensuring effective relation extraction. Extensive experiments on the Twitter dataset demonstrate that TGDI not only surpasses state-of-the-art baselines but also robustly suppresses the influence of irrelevant visual content in real-world multimodal settings.

Keywords: Multimodal relation extraction · Modality preference · Multimodal fusion · Information aggregate.

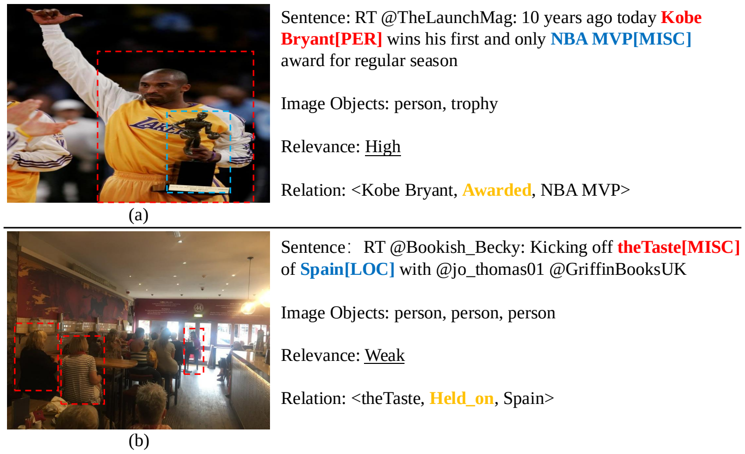
1 Introduction

In the era of big data, the vast amount of information available on the Internet presents significant challenges for data analysis and information extraction. Relation extraction is a crucial subtask within information extraction that effectively identifies structured data and valuable insights. This process advances search engine technology and enhances and enriches knowledge graphs. Text-based relation extraction has yielded excellent results in previous studies. Recently, there has been a substantial increase in text, visual, and audio content on social media. The vast surge in multimodal information benefits social media relation extraction, enabling public opinion monitoring, user relationship mining, and personalized content recommendations.

Therefore, multimodal relation extraction (MRE) has garnered increasing attention. MRE aims to classify the relationships between two given entities by utilizing visual information as supplementary evidence, thereby enhancing extraction performance. A notable characteristic of social media is the limited word count of text, resulting in concise messages that often provide few details and context. In contrast, visual content such as images and videos can provide essential background information, thereby enhancing the effectiveness of relation extraction. Previous work has extensively explored the alignment and fusion between images and text. The MEGA[25] model combines a parsed scene graph from the image with a syntactic dependency tree and textual semantics derived from the alignment of representations to learn mappings between visual and textual relationships. To fully leverage image information, most researchers use object images to align with the entity. Prior work has aimed to find correlations between entity-entity and object-object relationships [21], and explored alignment between entities and objects [6]. In addition, some studies [2][12] leverage visual prefixes, such as object or image features, to enhance multimodal fusion. How to establish in-depth interactions among complex modality-specific semantic information and bridge the gap between modalities remains a long-term challenge that requires continuous efforts.

Another often-overlooked phenomenon is that many studies are based on the assumption that images and text are aligned. However, this is not always the case. Research data shows that images and text do not always correspond perfectly, with only 15.8% of images providing meaningful context for tweets, while text is often embedded within the images [15]. In Fig.1, if the image and tweet are not highly relevant, visuals cannot provide rich supplementary information. Furthermore, in multimodal named entity recognition [7][19], it has been observed that the performance of multimodal models is not always better than that of unimodal approaches. Researches in multimodal sentiment analysis[8][27] also indicate that the contributions of different modalities can vary significantly. While previous works have demonstrated good performance, MRE models often assign equal importance to text and images, even though different modalities should hold varying significance in social media. When the provided entity pairs are textual, excessive reliance on visual information can introduce visual bias.

Inspired by the above, this paper introduces an innovative Text-Guided Dual Interaction (TGDI) model designed for multimodal relation extraction in social media. The TGDI model focuses on filtering out unnecessary visual information, leveraging text guidance to enhance multimodal relation extraction. To achieve this, the Cross-Modal Interaction module conducts a global level fusion for initial alignment. Meanwhile, the Text-Oriented Interaction module refines this process by preserving key visual information guided by the text. Additionally, the Text Modulated Matching Gate plays a crucial role in regulating the contributions of visual data and assessing image-text similarity to minimize any visual



Sentence: RT @TheLaunchMag: 10 years ago today **Kobe Bryant[PER]** wins his first and only **NBA MVP[MISC]** award for regular season

Image Objects: person, trophy

Relevance: High

Relation: <Kobe Bryant, **Awarded**, NBA MVP>

(a)

Sentence： RT @Bookish_Becky: Kicking off **theTaste[MISC]** of **Spain[LOC]** with @jo_thomas01 @GriffinBooksUK

Image Objects: person, person, person

Relevance: Weak

Relation: <theTaste, **Held_on**, Spain>

(b)

Fig. 1. On Twitter, the relevance between text and images may vary. Here are two different scenarios: (a) High text-image relevance. When the text and image are highly relevant, visual information can serve as an auxiliary cue to aid relation extraction. In this case, the visual objects "a person and a trophy" correspond to the textual entities "Kobe Bryant" and "NBA MVP," facilitating the extraction of the textual relation Awarded. (b) Low text-image relevance. The attached image simply shows a group of people dining at a restaurant, lacking any explicit Spanish food or cultural elements. The visual objects mainly consist of "person," which provides little help in identifying the "Held_on" relation between "theTaste" and "Spain."

noise. Finally, the fusion function is adaptable to various text-image scenarios, paving the way for effective relation extraction. In summary, here are the main contributions of our paper:

- We propose a Text-Guided Dual Interaction (TGDI) model that takes text as the dominant modality and introduces a dual-interaction mechanism to capture fine-grained cross-modal features, adapting to varying degrees of text-image relevance.
- We design a Text Modulated Matching Gate to filter out irrelevant visual content, improving the alignment precision of multimodal relation extraction.
- We conduct extensive experiments on the Twitter dataset, demonstrating the effectiveness of TGDI and highlighting the importance of text dominance in MRE for social media contexts.

2 Related work

Multimodal relation extraction is a key branch of information extraction that has garnered increasing attention in recent years. Particularly in social media, researchers have discovered that visual information can supplement the semantic details missing from the text, leading to significant improvements in extraction performance[26]. Chen et al.[2] designed a hierarchical multi-scaled visual representation as visual guidance for fusion, utilizing both object images and whole images as visual prefixes for each self-attention layer in BERT. Dai et al.[3] developed an image-text matching approach to enhance the model's ability to capture different semantic correspondences by constructing hard negatives for improvement. Li et al.[12] proposed two types of prefix tuning, entity-oriented prefix and object-oriented prefix, to integrate deeper associations between intra-modal and inter-modal data. They also designed a dual-gated fusion module that identifies and suppresses irrelevant interaction data through local and global visual contexts. Hu et al.[6] introduced a pretraining task for entity-object and relation-image alignment, extracting self-supervised signals from large numbers of unlabeled image-caption pairs and providing soft pseudo-labels to guide the pretraining process. Wu et al.[16] utilized visual and textual scene graph structures to represent input data, integrating them into a cross-modal graph. They refined the structure guided by the graph information bottleneck principle and introduced latent multimodal topic features to tackle the challenges of internal and external information utilization in multimodal relation extraction. Yuan et al.[21] identified correlations between object-object and entity-entity relationships and introduced an edge-enhanced graph alignment network that aligns nodes and edges across graphs to improve joint multimodal entity-relation extraction. Xu et al.[17] proposed a reinforcement learning-based data segmentation approach to determine whether social media posts are better suited for multimodal or unimodal models. Shen et al.[13] argued that previous methods overlooked textual information within images, resulting in performance degradation when handling text-intensive images. They incorporated cross-attention in the textual evidence integration process to extract entity-related information from image captions and OCR text.

Unlike the above studies, we treat different modalities unequally. We aim to explore modality preferences and guide modal interaction through the dominant textual modality. Additionally, we aim to correct the bias introduced by visual information.

3 methodology

3.1 Problem Definition

Given a short sentence $T = \{t_1, t_2, \ldots, t_m\}$ that contains an entity pair (e_s, e_o) and associates with an image I. The main objective of MRE is to identify the semantic relationship $r \in R = \{r_1, r_2, \ldots, r_n\}$ between the entities. The process is defined as a function $F : (e_s, e_o, T, I) \rightarrow R$.

The overall architecture of TGDI for multimodal relation extraction in social media is illustrated in Fig.2, with a detailed explanation of each module provided in the following section.

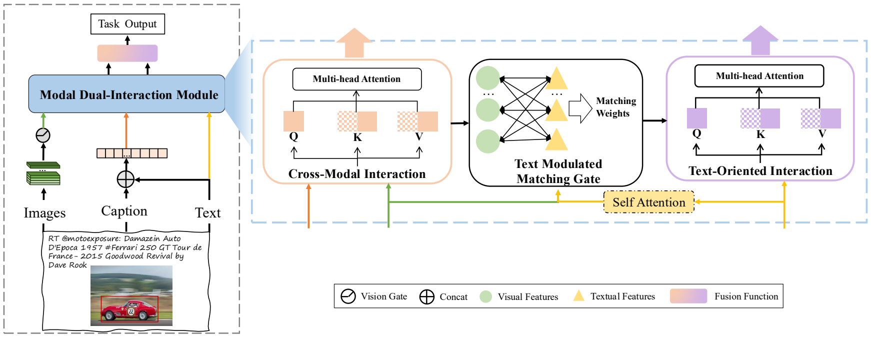

Task Output
Modal Dual-Interaction Module
Images
Caption
Text
RT @motoexposure: Damazein Auto D'Epoca 1957 #Ferrari 250 GT Tour de France - 2015 Goodwood Revival by Dave Rook
Multi-head Attention
Q
K
V
Cross-Modal Interaction
Matching Weights
Text Modulated Matching Gate
Multi-head Attention
Q
K
V
Text-Oriented Interaction
Self Attention
Vision Gate
Concat
Visual Features
Textual Features
Fusion Function

Fig. 2. The overall framework of our proposed TGDI introduces a modal dual-interaction module designed to minimize the interference of visual noise while preserving highly relevant fused features for textual relations. Additionally, we employ a result fusion function to accommodate different text-image relevance scenarios.

3.2 Feature extraction

Visual representation Global image features mainly capture comprehensive information in the image, including abstract concepts such as scenes and themes. The objects typically refer to a specific entity and serve as key features in an image, making them well-aligned with textual entities. The object image $O = (o_1, o_2, \ldots, o_m)$ is obtained by extracting m salient objects from the entire image using the Visual Grounding Toolkit[18]. Combining both global and object image features can effectively leverage the potential value of visual modality. The vision transformer(ViT)[5] as the visual encoder, global image I and object image O are rescaled to 224×224 pixels and fed to ViT to obtain the hierarchical visual semantic representation as $V = \{V_0, V_1, \ldots, V_{12}\}, V_i \in \mathbb{R}^{n \times d}$. To ensure that the visual features obtained align with the structure of the textual information, we apply a mapping function to transform the visual features accordingly.

$$V_i = Conv_{(1\times1)}(MLP(Pool(V))), \tag{1}$$

where $Pool$ is a global average pooling layer that reduces the dimensions of visual features, then a 1×1 kernel size convolution module to align with textual features. The dynamic vision gate function generates i-th layer probability vector $g(l)$, which is applied to hierarchical global visual features to obtain the final visual features that match the i-th layers in the fusion model. Then we allocate visual information from different hierarchies to specific layers, enabling the effective fusion of information to enhance the model's text and image fusion capabilities. The vision gate is calculated as follows:

$$s^{(l)} = L(W_l(\frac{1}{c}\sum_{i=1}^{c} P(V_i))), \tag{2}$$

$$g^{(l)} = softmax(s^{(l)}), \tag{3}$$

where l denotes the index of each layer, L is activate function Leaky_ReLU, P denotes pooling layer. Then the final aggregated hierarchical visual feature V obtained with $g^{(l)}$ to match the l-th layer $V = g^{(l)}V^{(l)}$. Global visual features and object features are fused together as $\tilde{V}$ for subsequent steps.

Text representation Image captions serve as an essential source of information, helping models better understand the content of images and their relationship with text. Visual context can be directly integrated by utilizing captions to expand and enhance the original text, thereby reducing the semantic gap caused by differences in modalities across various spaces. Caption (C) generated by the image-to-text generation model of Bilp2[9] extends the textual content and integrates comprehensive semantic information from the image, thereby offering rich background evidence for text. For text modality, we use BERT[4] as an encoder to obtain text representations. Each input includes text and caption as textual modal information $T_e = \{t_0, t_1, t_2, \ldots, t_m, t_{m+1}, t_{c1}, \ldots, t_{cn}, t_{cn+1}\}$, where t_0 represent begin token [CLS], t_{m+1} and t_{cn+1} represent end tokens [SEP], t_1 to t_m as the textual sequence, t_{c1} to t_{cn} as the caption sequence. Therefore, the enhanced textual representation of the input is obtained through BERT: $H_e = BERT(T_e)$. Similarly, the input text T is processed through BERT to obtain its representation $H = BERT(T)$.

3.3 Modal Dual-Interaction module

To better align and integrate images and text, we propose a Modal Dual-Interaction module. Specifically, visual and textual information are fused at a global level through Cross-Modal Interaction (CMI). Subsequently, the Text Modulated Matching Gate computes matching weights based on both visual and textual features. The refined features, after mitigating visual bias, are fed into the Text-Oriented Interaction (TOI) module to generate well-aligned multimodal representations for relation extraction.

Cross-Modal Interaction In multimodal fusion, hierarchical visual features are used as visual prompts appended with the textual sequence at each BERT attention layer, aiming to guide the fusion with layer-level textual representations. For input sequence T_e, the context representation of $H_e^{l-1} \in \mathbb{R}^{n \times d}$:

$$Q^l = H_e^{l-1} W_Q^l, K^l = H_e^{l-1} W_K^l, V^l = H_e^{l-1} W_V^l, \tag{4}$$

W_Q^l, W_K^l, W_V^l are mapping parameters of attention. For the l-th visual representation, we use a linear transformation W_l^{α} to project $\tilde{V}^(l)$ into the same embedded space as the text representation to get the visual prompt vector:

$$\{\alpha_l^k, \alpha_l^v\} = \tilde{V}_l W_l^{\alpha}. \tag{5}$$

The hidden features at the l-th layer of the fusion encoder based on CMI are calculated as:

$$f_1^l = softmax(\frac{Q^l[\alpha_l^k; K^l]^\top}{\sqrt{d}})[\alpha_l^v; V^l]. \tag{6}$$

Text Modulated Matching Gate The TMMG module enables a finer-grained computation of the similarity between each patch in an image and each token in the text. This module captures more precise local cross-modal alignment, enhancing the correspondence between modalities while reducing visual noise, thereby improving the accuracy of cross-modal matching. The textual representation H is then fed into a self-attention mechanism to capture the dependencies and syntactic relationships within the text, resulting in a deeper understanding of the textual information.

$$H_s = softmax(\frac{HW_Q(HW_K)^\top}{\sqrt{d}})HW_V, \tag{7}$$

where W_Q, W_K, and $W_V \in \mathbb{R}^{d\times d}$ are query, key, and value trainable weight matrices, respectively.

To minimize the impact of noise from irrelevant images, we regulate the contributions of visual information through a matching gate mechanism that assesses the matching degree between the image and the text. The Text Modulated Matching Gate between textual features $H_s \in \mathbb{R}^{b\times l\times d}$ and visual features $V \in \mathbb{R}^{b\times n\times d}$ is calculated as:

$$A = softmax(VH_s), \tag{8}$$

$$C' = AH_s, \tag{9}$$

$$m = M\left(\frac{1}{N}\sum_{i=1}^{N}[V; C'; V \odot C']\right), \tag{10}$$

where $\odot$ represents Hadamard product, $M(\cdot)$ is an MLP. Based on the similarity gate, we can obtain the final visual-aware fusion representations:

$$f_2 = m \odot f_1. \tag{11}$$

Tex-Oriented Interaction In TOI, the textual features interact with the fusion representation obtained from above module, allowing the text to rectify any inaccurate fusion information. This process ultimately results in a text representation that effectively incorporates relevant visual content. For text representation of $H_s^{l-1} \in \mathbb{R}^{n\times d}$:

$$Q^l = H_s^{l-1}W_Q^l, K^l = H_s^{l-1}W_K^l, V^l = H_s^{l-1}W_V^l. \tag{12}$$

Then, f_2 is reshaped and split into key-value prefixes $\left\{\beta_l^k, \beta_l^v\right\}$ of each layer, enabling layer-specific injection of enhanced fusion context into the attention computation. And the hidden state at $l-$th layer of TOI is calculated as follows:

$$f_3^l = softmax(\frac{Q^l[\beta_l^k; K^l]^\top}{\sqrt{d}})[\beta_l^v; V^l]. \tag{13}$$

3.4 Result fusion and classification

Considering the varying amounts of information contained in the two interaction results, we design a weighted fusion function to merge them, yielding a more reliable output. The output f_3 from TOI is the primary foundation for adjustment. The core computational process can be stated as follows:

$$\mathbf{R} = \lambda \cdot f_3 + (1 - \lambda) \cdot f_1, \tag{14}$$

where $\lambda \in [0, 1]$. MRE aims to predict r when the input consists of a post T and an image I. Ultimately, the MRE loss equation is given below, where W and b are learnable parameters.

$$p(r|e_o, e_s, T) = softmax(W\mathbf{R} + b), \tag{15}$$

$$\mathcal{L} = -log(p(r|e_o, e_s, T)). \tag{16}$$

4 Experiments

4.1 Experiment settings

Dataset and evaluation metrics Regarding the dataset, we utilized MNRE[4][25] (see in Table 1) from Twitter to experiment, which comprises 9201 images and 30970 entities across 23 relations, with a split of 12247/1624/1614 samples for the train/validation/test sets. The metrics used in experiments include Accuracy (Acc%), Precision (P%), Recall (R%), and F1 score (F1%).

Table 1. The statistics of MNER dataset

Dataset	#Image	#Sentence	#Instance	#Entity	#Relation
MNRE	9201	9201	15485	30970	23

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$P = \frac{TP}{TP + FP} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{20}$$

[4] The dataset is available at `https://github.com/thecharm/MNRE`.

Experiments Setup Our model, which utilizes PyTorch 2.0.0, is trained on a NVIDIA RTX A6000 GPU. All input images are resized to a standard resolution of 224×224 pixels. Both visual and textual modalities are transformed into 768-dimensional hidden representations through CLIP-ViT-B/32 and BERT-base-uncased, respectively. We employ the AdamW optimizer with a learning rate of 3e-5 and a batch size of 16. The number of object images (m) is set to 3. The hyperparameter λ is set to 0.5 to balance the fusion strategy. To mitigate overfitting, we apply a dropout rate of 0.01 during training.

Baselines We compare our model with well-known MRE baselines as follows:

- **Text Baseline:**
  1. Glove + CNN [23]: A CNN-based model for relation extraction.
  2. PCNN [22]: A distantly supervised model that uses external knowledge graphs to assign labels to sentences containing the same entities automatically.
  3. MTB [14]: A relation extraction model that enhances BERT pre-training by incorporating entity masking.
- **Multimodal Baseline:**
  1. VisualBERT [11]: A pre-trained visual-language model for capturing rich semantics in images and associated text.
  2. UMT [20]: A unified multimodal transformer integrating a multimodal interaction module and an entity span detection module for prediction.
  3. UMGF [24]: A unified multimodal graph framework capturing multimodal semantic relationships via graph-based fusion layers.
  4. BERT + SG [25]: Combines fine-tuned BERT representations with visual features extracted from a pre-trained scene graph (SG).
  5. BERT + SG + Att [25]: Incorporates an attention mechanism to compute semantic similarity between visual graphs and textual contents.
  6. MEGA [25]: A multimodal neural network with efficient graph alignment, considering the structural similarity and semantic agreement between visual and textual graphs.
  7. IFAformer [10]: A dual-transformer architecture introducing a visual prefix for modal fusion to reduce sensitivity to errors.
  8. HVPNeT [2]: A hierarchical visual prefix fusion network leveraging hierarchical visual features to alleviate error sensitivity from irrelevant images.
  9. MKGformer [1]: Employs a correlation-aware fusion module to mitigate the impact of noisy information.

4.2 Overall Results

Main Results The experimental results demonstrate that our model outperforms the baseline across all metrics, validating the effectiveness of our approach, as shown in Table 2. Notably, the following points deserve attention: First, not all multimodal methods outperform unimodal methods; however, it is undeniable that images can effectively enhance certain context information in social media. Second, while the MEGA model attempts to leverage image and text alignment through parsed scene graphs, we observed that irrelevant images often introduce visual noise. TGDI enhances performance by minimizing noise and facilitating deep interaction between the two modalities. Finally, compared to models HVPNeT, MKGformer, and IFAformer, which also focus on reducing visual error sensitivity, TGDI prioritizes text-oriented relational extraction while decreasing visual bias. This strategy enhances performance and highlights the critical role of text modality in extraction within social media contexts.

Table 2. Performance comparison of previous SOTA baseline models for multimodal RE on MNRE dataset

Modal	**Methods**	**Precision**	**Recall**	**F1**
Text	Glove + CNN	57.81	46.25	51.39
	PCNN	62.85	49.69	55.49
	MTB	64.46	57.81	60.86
Multimodal	VisualBERT	57.15	59.48	58.30
	BERT + SG	62.95	62.65	62.80
	BERT + SG + Att	60.97	66.56	63.64
	UMT	62.93	63.88	63.46
	UMGF	64.38	66.23	65.29
	MEGA	64.51	68.44	66.41
	IFAformer	82.59	80.78	81.67
	HVPNeT	83.64	80.78	81.85
	MKGformer	82.67	81.25	81.95
	TGDI	**84.18**	**83.13**	**83.65**
Ablation	*w/o* caption	83.86	82.81	83.33
	w/o TMMG	83.36	83.75	83.55
	w/o CMI	81.75	81.87	81.81
	w/o TOI	82.28	81.25	81.76
	w/o R	81.75	81.88	81.81

At the same time, Fig.3 compares extraction accuracy across several models, clearly indicating that TGDI achieves the highest accuracy. Unlike the coarse-grained fusion between modalities, our model fully leverages modality information through deep interactions and text-oriented corrections. It strategically integrates representations from the two interaction modules, considering the extraction results under varying scenarios.

Ablation study To assess the contribution of each module to MRE, we conducted an ablation study using consistent parameters to confirm their effectiveness. Specifically, we removed the following components: "image caption," "Text Modulated Matching Gate," "Cross-Modal Interaction," "Text-Oriented Interaction," and "Fusion Function," represented as "w/o caption," "w/o TMMG," "w/o CMI," "w/o TOI," and "w/o R". The detailed results are shown in Table 2.

After removing the caption, Precision, Recall, and F1 each decreased by 0.32%. This suggests that captions are a direct means of incorporating visual information into the textual modality. By doing so, they help bridge the semantic gap between modalities, reducing the impact of modality differences on performance. The decline in Precision and F1 for the TMMG highlights its primary function of assessing the relevance between images and texts. Without CMI in result fusion, all metrics show a noticeable decline, indicating that visual information is an important component that provides supplementary visual cues. In removing the TOI, Recall decreased by 1.88% and F1 by 1.89%, demonstrating that global level fusion of image and text yields poor results. Regarding the result fusion function R, using the TOI output representation for extraction directly results in a decrease in Precision by 2.75%, Recall by 1.25%, and F1 by 1.84%. This decline suggests that integrating results from different interaction levels helps combine perspectives from relevant and irrelevant scenarios, leading to more stable and robust outcomes.

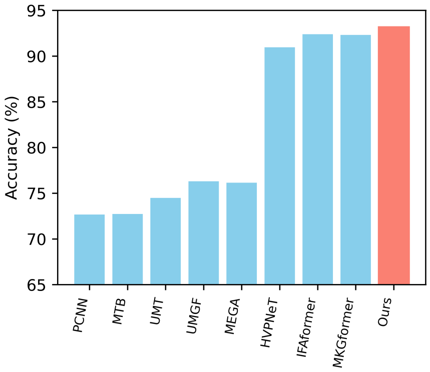


Fig. 3. Comparison extraction accuracy of different models

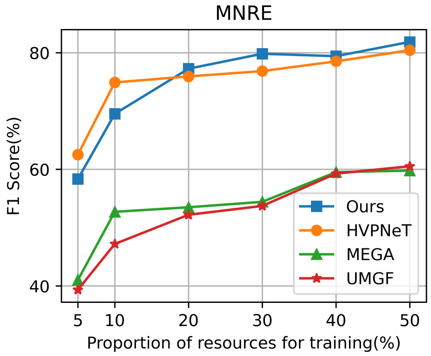


Fig. 4. Performances on low-resource setting on MNER

Analysis Based on our analysis, we conclude that deep interaction and fine-grained fusion between text and images are essential, confirming the dominant role of the text modality. Captions enhance the contextual information of images, reduce ambiguity, and minimize visual bias. Furthermore, the lack of sufficient interaction between images and text necessitates a more profound fusion. TOI can effectively improve this interaction, align text with image information, and demonstrate its effectiveness in a text-oriented manner. Reducing visual noise minimizes visual bias, leading to more accurate fusion outcomes. Because irrelevant images and text are often shared on social media, the TMMG helps balance these relationships in multimodal fusion, enabling better control of visual information and reducing excessive noise, which enhances detail in interactions. Regarding result fusion, the decrease in values indicates that the fusion of results from two different interaction levels can combine outcomes from both learning phases. It ensures that suitable image information is fully fused while preventing mismatched image-text pairs from interfering with the extraction results, thus avoiding unstable interaction fluctuations and resulting in more stable and robust outcomes.

4.3 Case study and discussion

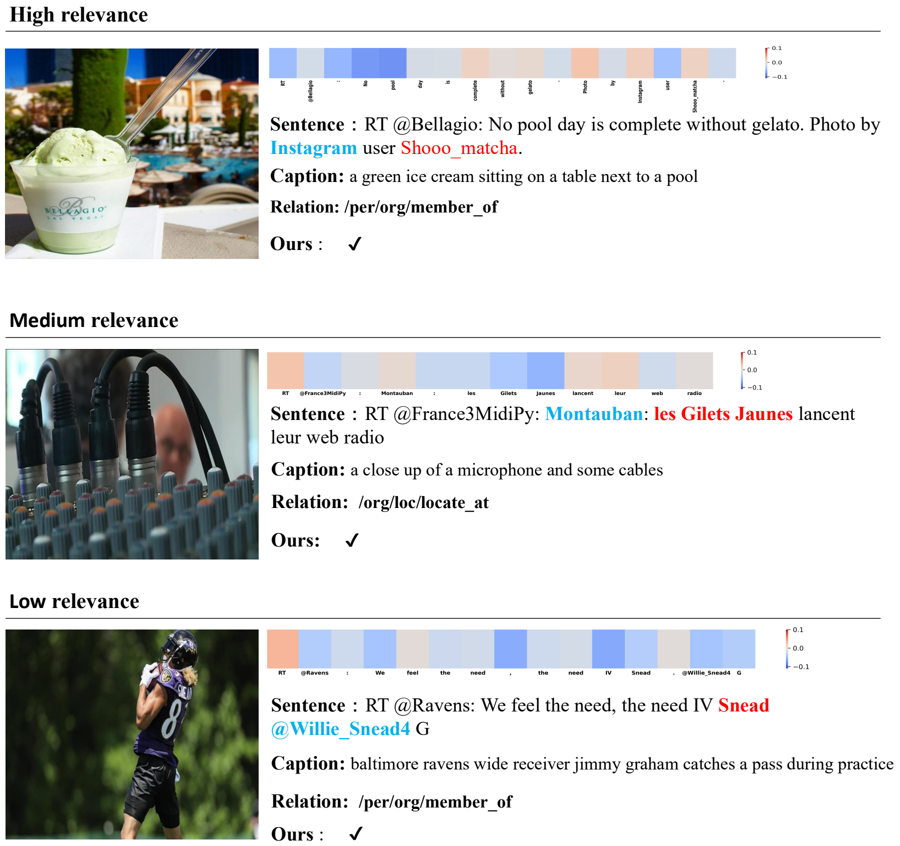
Performance on Categories Compared with HVPNeT, our model achieves significant performance improvements in most of the six main categories of the MNRE test set (Table 3). It effectively integrates both text and visual content. For categories like "person," where clear objects are present, the text aligns well with the images, allowing relevant visual information to aid in relation extraction. In contrast, for categories such as "location," "organization," and certain "misc" types that do not correspond to regional images, our model relies more on textual information to make accurate judgments.

Table 3. Our results compared with HVPNeT on the six main categories of the test set (HVPNeT results are based on our reproduction).

Category	**Count**	**HVPNeT(F1.)**	**TGDI(F1.)**
`/per/per/peer`	156	88.96	91.69
`/per/org/member_of`	110	80.33	83.70
`/loc/loc/contain`	99	93.88	96.55
`/per/misc/present_in`	74	75.00	83.69
`/org/loc/locate_at`	46	85.06	83.72
`/per/loc/place_of_residence`	29	63.16	64.41

Low resource We conducted experiments in low-resource settings, randomly sampling 5% to 50% from the original training set in a low-resource scenario. Fig.4 shows a comparison of our model with several baselines. It is clear that in most low-resource scenarios, reducing visual noise can effectively mitigate bias and yield better results. However, when the training set is particularly small, excessive reliance on textual information can lead to a significant loss of contextual background, resulting in poorer performance.

Case study To validate our model, we selected several relevant image-text pairs for analysis in the MNRE task (Fig.5). In these image-text posts, excessive visual information can mislead the model, resulting in incorrect predictions and interfering with judgment. Building on our previous analyses, this finding reinforces that abundant visual information does not always lead to better model performance; instead, it can introduce noise and hinder accurate predictions. This highlights the importance of our proposed text-guided approach, prioritizing textual information to effectively integrate visual cues, ensuring a more reliable and meaningful multimodal representation.



High relevance

Sentence : RT @Bellagio: No pool day is complete without gelato. Photo by **Instagram** user Shooo_matcha.

Caption: a green ice cream sitting on a table next to a pool

Relation: /per/org/member_of

Ours : ✓

Medium relevance

Sentence : RT @France3MidiPy: **Montauban**: **les Gilets Jaunes** lancent leur web radio

Caption: a close up of a microphone and some cables

Relation: /org/loc/locate_at

Ours: ✓

Low relevance

Sentence : RT @Ravens: We feel the need, the need IV **Snead** **@Willie_Snead4** G

Caption: baltimore ravens wide receiver jimmy graham catches a pass during practice

Relation: /per/org/member_of

Ours : ✓

Fig. 5. A selection of image-text pairs with varying relevance is analyzed, including images and text from Twitter, a similarity heatmap between the text and the image, a caption generated from the image, and predefined relationship labels between entities. (Red represents the head entity, while blue represents the tail entity.)

5 Conclusion

In this paper, we propose the Text-Guided Dual Interaction Multimodal Relation Extraction (TGDI) model to address the weak correlation between text and images in social media relation extraction. At its core, TGDI employs a Modal Dual-Interaction mechanism. The interaction module consists of a Cross-Modal Interaction, which performs global fusion of text and image features, and a Text-Oriented Interaction, which eliminates irrelevant visual information, reduces visual noise, and enhances the relevance of visually perceptive text representations. And the Text Modulated Matching Gate regulates visual contributions and evaluates image-text similarity to minimize visual noise. To effectively integrate the outcomes of both modules, we introduce a fusion function that balances predictions based on image-text alignment and those relying solely on textual information. Extensive experiments demonstrate that TGDI significantly improves relation extraction performance, highlighting the effectiveness of our approach.

6 Limitations

Our model focuses on low text-image relevance and is text-dominant, evaluated on a single benchmark dataset. Its performance may be restricted in multimodal applications where text and image correlation is high, and visual information supplements text. We will explore this in future work.

References

1. Chen, X., Zhang, N., Li, L., Deng, S., Tan, C., Xu, C., Huang, F., Si, L., Chen, H.: 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. pp. 904–915 (2022)
2. Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., Huang, F., Si, L., Chen, H.: 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. pp. 1607–1618. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.121
3. Dai, Y., Gao, F., Zeng, D.: An alignment and matching network with hierarchical visual features for multimodal named entity and relation extraction. In: International Conference on Neural Information Processing. pp. 298–310. Springer (2023)
4. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hu, X., Chen, J., Liu, A., Meng, S., Wen, L., Yu, P.S.: Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5185–5194 (2023)

7. Jia, M., Shen, L., Shen, X., Liao, L., Chen, M., He, X., Chen, Z., Li, J.: 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. pp. 8032–8040 (2023)
8. Lei, Y., Yang, D., Li, M., Wang, S., Chen, J., Zhang, L.: Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences. In: CAAI International Conference on Artificial Intelligence. pp. 189–200. Springer (2023)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
10. Li, L., Chen, X., Qiao, S., Xiong, F., Chen, H., Zhang, N.: On analyzing the role of image for visual-enhanced relation extraction (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 16254–16255 (2023)
11. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
12. Li, Q., Guo, S., Ji, C., Peng, X., Cui, S., Li, J.: Dual-gated fusion with prefix-tuning for multi-modal relation extraction. arXiv preprint arXiv:2306.11020 (2023)
13. Shen, Q., Lin, H., Liu, H., Lin, Z., Wang, W.: Watch and read! a visual relation-aware and textual evidence enhanced model for multimodal relation extraction. In: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). pp. 2491–2496 (2024)
14. Soares, L.B., FitzGerald, N., Ling, J., Kwiatkowski, T.: Matching the blanks: Distributional similarity for relation learning. arXiv preprint arXiv:1906.03158 (2019)
15. Vempala, A., Preoţiuc-Pietro, D.: 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. pp. 2830–2840 (2019)
16. Wu, S., Fei, H., Cao, Y., Bing, L., Chua, T.S.: Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. arXiv preprint arXiv:2305.11719 (2023)
17. Xu, B., Huang, S., Du, M., Wang, H., Song, H., Sha, C., Xiao, Y.: 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. pp. 1855–1864 (2022)
18. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4683–4693 (2019)
19. Yu, J., Jiang, J., Yang, L., Xia, R.: 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. pp. 3342–3352. Association for Computational Linguistics (2020)
20. YU, J., JIANG, J., YANG, L., XIA, R.: 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. pp. 3342–3352 (2020)
21. Yuan, L., Cai, Y., Wang, J., Li, Q.: Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In: Proceedings of the AAAI conference on artificial intelligence. pp. 11051–11059 (2023)

22. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1753–1762 (2015)
23. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. pp. 2335–2344 (2014)
24. 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. pp. 14347–14355 (2021)
25. Zheng, C., Feng, J., Fu, Z., Cai, Y., Li, Q., Wang, T.: Multimodal relation extraction with efficient graph alignment. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 5298–5306 (2021)
26. Zheng, C., Wu, Z., Feng, J., Fu, Z., Cai, Y.: 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). pp. 1–6 (2021). https://doi.org/10.1109/ICME51207.2021.9428274
27. Zhou, M., Quan, W., Zhou, Z., Wang, K., Wang, T., Yan, D.M.: Tcan: Text-oriented cross attention network for multimodal sentiment analysis. arXiv preprint arXiv:2404.04545 (2024)