Dynamic Visual Evidence Reliability Modeling for Multi-Image Multimodal Named Entity Recognition

SONG Lizhi, XU Mingying, LIU Jie*

the School of Artificial Intelligence and Computer Science, North China University of Technology, BeiJing 100144, China

Abstract: Multi-image multimodal named entity recognition (MNER) not only requires the exploitation of richer visual context, but also faces a more fundamental challenge: different images contribute unequally to entity recognition. Existing multi-image MNER methods typically either aggregate multiple images directly or assume that different images and modalities make approximately stable contributions during fusion. Such static treatment can easily amplify noise in the presence of weakly relevant or redundant images. To address this issue, this paper reconceptualizes multi-image MNER as a task of dynamic visual evidence reliability modeling and proposes a visual evidence reliability modeling framework. Specifically, the framework first encodes multiple images at the image level and constructs a joint visual representation through an inter-image interaction module. It then estimates the visual evidence reliability score based on the consistency between textual and visual representations. Finally, a regulation module consisting of multiple progressive reliability regulation units is introduced to calibrate the contribution of the visual branch in a layer-wise manner, thereby following a simple principle: first assess whether the visual evidence is reliable, and then determine the extent to which it should participate in fusion. Unlike one-shot static fusion, the proposed method can continuously adjust the influence of visual information throughout multi-layer cross-modal interactions, making it better suited to multi-image scenarios where visual support is inherently unstable. Experimental results show that the proposed method achieves competitive performance on the multi-image MNER benchmark. By re-examining the fusion process in multi-image MNER from the perspective of visual evidence reliability, this paper offers a more direct modeling paradigm for multi-image multimodal sequence labeling.

Keywords: Multi-image multimodal named entity recognition; Visual evidence reliability modeling; Multimodal fusion; Cross-modal interaction; Sequence labeling

I. Introduction

Named entity recognition (NER) is a fundamental task in information extraction and has been widely applied in knowledge graph construction, question answering, and social media analysis. As social media content becomes increasingly multimodal, multimodal named entity recognition (MNER), which incorporates visual information, has emerged as an important direction for alleviating semantic sparsity in short texts. Compared with text-only models, images can provide complementary context for both entity boundary detection and entity type classification, especially when textual expressions are brief,

* *Corresponding author*

colloquial, or ambiguous. In recent years, research in this area has gradually extended from single-image settings to multi-image settings, and the introduction of datasets such as MNER-MI and MNER-MI-Plus, has provided a clearer benchmark foundation for studying multi-image MNER.

However, the challenge in multi-image settings goes beyond the mere increase in the number of input images. Compared with single-image input, multiple images introduce not only richer context but also a more complicated distribution of evidence: some images contain visual cues that are directly relevant to entity recognition, some provide only weakly related background information, and others may even introduce noise. Existing studies have shown that many multi-image methods still incorporate multiple images through direct averaging, simple aggregation, or uniform processing. Such static fusion strategies can easily weaken the discriminative ability of the model when noisy images are present. As discussed in prior work, existing methods often assign equal weights to images or modalities, even though image-text relevance varies across samples. Inappropriate weight assignment may therefore cause the model to overlook critical information or amplify irrelevant noise.

Based on this observation, we argue that the core issue in multi-image MNER is not how to additionally select a particular category of visual objects, but how to determine whether the visual evidence is reliable for a given sample and how strongly it should contribute to entity recognition. If a model cannot explicitly distinguish strong evidence from weak evidence, multi-image input may shift from being a source of complementary information to a source of interference. This issue is particularly evident in cases where the text already provides sufficient information and some images serve only as peripheral context, in which case uniform fusion often lacks selectivity. Conversely, when the text is ambiguous and visual cues become crucial, static fusion may fail to fully exploit the value of image information. These observations suggest that simply adding more visual input does not automatically yield effective gains in multi-image MNER; the key lies in whether the model can regulate the use of visual information according to sample-specific differences.

To address this issue, this paper reformulates the cross-modal fusion process in multi-image MNER as a problem of dynamic visual evidence reliability modeling and proposes a visual evidence reliability modeling framework. Specifically, the proposed framework first encodes multiple images at the image level and obtains a joint visual representation through an inter-image interaction module. It then estimates a visual evidence reliability score based on the consistency between textual and visual representations, which characterizes how strongly the visual modality supports the current entity recognition task. Finally, through a regulation module composed of multiple progressive reliability regulation units, the framework continuously calibrates the influence of the visual branch on multimodal representation updating during cross-modal interaction. Rather than injecting visual features into the text only once, this design repeatedly adjusts the strategy of whether visual information should be more heavily relied upon across multiple fusion layers.

The main contributions of this paper are as follows:

This paper re-examines multi-image MNER from the perspective of unstable visual evidence contribution and formulates it as a problem of dynamic visual evidence reliability modeling.

This paper proposes a fusion framework

consisting of a visual evidence reliability estimator and a progressive reliability regulation module, which adaptively adjusts the extent to which visual information contributes to entity recognition according to sample-specific differences.

This paper demonstrates the effectiveness of the proposed framework on multi-image MNER benchmarks.

II. Related Work

Named Entity Recognition (NER) is a foundational task in information extraction. Early research focused on text-only NER, transitioning from statistical models like conditional random fields (CRFs) to deep learning architectures such as BiLSTM-CRF and Transformer-based models (e.g., BERT). To address semantic sparsity and ambiguity in short social media texts, single-image MNER emerged to leverage visual context for improving entity boundary detection and classification. These methods typically employ co-attention mechanisms or unified multimodal transformers to align textual and visual features. However, single-image models often falter when the visual cue is noisy or when a post contains multiple images providing either complementary or conflicting information.

Consequently, the research landscape has shifted toward multi-image scenarios. Huang et al. pioneered this direction by introducing MNER-MI and MNER-MI-Plus, the first large-scale benchmarks for multi-image MNER, and proposed a temporal prompt model to handle sequential visual inputs. Despite this progress, many existing multi-image methods still rely on static aggregation strategies, such as direct averaging, simple concatenation, or uniform sequential modeling. As noted in recent studies, these “one size-fits-all” fusion approaches fail to distinguish between highly relevant visual evidence and peripheral noise, leading to semantic interference where irrelevant images degrade the performance of the textual backbone. To mitigate visual noise, researchers have explored more flexible fusion paradigms, including attention-based weighting, gating mechanisms, Mixture-of-Experts (MoE), and prompt-based fusion. For instance, Zhao et al. proposed a graph interaction framework based on text-image relevance, which models fine-grained correlation between text and multiple images to filter irrelevant visual information for MNER, which utilizes reinforcement learning and heterogeneous graph transformers to filter irrelevant images. Wu and Liu introduced Symmetric Fusion with contrastive learning to bridge the modality gap in an exclusive space, while Hu et al. developed VS-MRC to select meaningful visual representations through an aggregation-mapping-comprehension paradigm.

While these methods improve relevance filtering, they often treat visual reliability as a binary selection or a static gate. In contrast, our work moves beyond simple filtering by drawing on contrastive learning to model the consistency between cross-modal representations. We propose a dynamic visual evidence reliability modeling approach that estimates a sample-level reliability score based on text-visual alignment. Unlike prior works that inject visual features in a single pass, our framework employs multi-layer progressive reliability regulation units. This design allows the model to continuously calibrate the influence of the visual branch throughout the deep interaction process, ensuring that visual information is heavily utilized only when confirmed as reliable evidence. This effectively prevents interference in text sufficient or noisy scenarios. Furthermore, inspired by recent advances in multimodal large language models and graph-based modeling, our framework seeks to refine cross-modal interaction through a

more granular, evidence driven perspective.

III. Method

A. Task definition

Given an input text sequence $S = \{s_1, s_2, \ldots, s_n\}$ and its corresponding image set $I = \{I_1, I_2, \ldots, I_m\}$, the multimodal named entity recognition task in Multi-image scenarios aims to identify named entities from the text and assign a predefined label to each token. In this paper, we formulate the task as a sequence labeling problem, where the goal is to predict the corresponding BIO label sequence

$$Y = \{y_1, y_2, \ldots, y_n\}$$

where $y_i \in Y$ and Y denotes the label space.

Unlike methods that directly fuse uniformly aggregated image features, our approach explicitly models how strongly visual evidence supports the current sample. Specifically, the model not only learns a joint representation of textual and visual information, but also estimates the reliability of the visual modality for each sample and uses it to regulate the extent to which visual information contributes to entity recognition.

B. Overall framework

We propose a visual evidence reliability modeling framework consisting of four components: a Multi-image visual encoder, a text encoder, a visual evidence reliability estimator, and a progressive reliability regulation module. The overall framework of the VERC-MI is shown in Figure 1. First, the model encodes multiple input images and derives a joint visual representation through an inter-image interaction module. Meanwhile, a pretrained language model is employed to extract contextualized textual representations. Then, a visual evidence reliability score is estimated based on the consistency between textual and visual representations, which reflects how strongly the visual modality supports the current entity recognition task. Finally, within the progressive reliability regulation module, the model progressively calibrates the influence of the visual branch on multimodal representation updating through multi-layer cross-modal interactions, and the resulting representation is fed into a CRF decoder for label prediction.

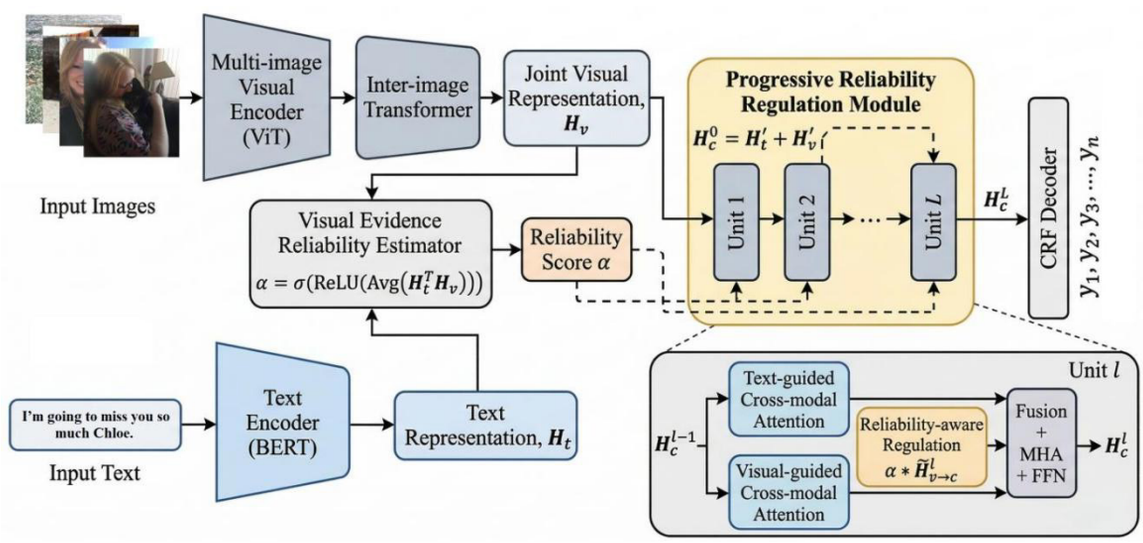


Figure 1 The overall framework of VERC-MI

The core idea of the framework is straightforward: it first evaluates whether the visual evidence is reliable for the current sample, and then determines how much it should contribute during fusion. In this way, the model can adaptively regulate the use of visual information according to

sample-specific image-text relationships.

C. Multi-image visual encoding

To obtain a unified Multi-image visual representation, we model the input images directly at the image level. For each input image I_j, its visual representation is first extracted by a Vision Transformer:

$$v_j = \mathrm{ViT}(I_j), \quad j \in \{1,2,\ldots,m\},$$

where $v_j \in \mathbb{R}^{d_v}$ and d_v denotes the visual feature dimension.

All image representations are then organized into an image sequence:

$$V = (v_1, v_2, \ldots, v_m) \in \mathbb{R}^{m \times d_v}.$$

Since different samples may contain different numbers of images, the image sequence is padded to a fixed length m. Based on this sequence, an inter-image interaction encoder is applied to capture dependencies among different images:

$$\tilde{V} = \mathrm{Transformer}(V),$$

where $\tilde{\mathbf{V}} \in \mathbb{R}^{m \times d_v}$ denotes the joint visual features after interimage interaction.

To obtain a sample-level visual representation, the joint visual features are further aggregated:

$$H_v = \mathrm{Pool}(\tilde{V}),$$

where $H_v \in \mathbb{R}^{d_v}$. We use average pooling to derive a compact Multi-image visual representation. After this process, the model obtains a sample-level visual representation that captures the overall semantics of multiple images, which serves as the input for subsequent reliability estimation and cross-modal regulation.

D. Text encoding

For the input text sequence $S = (s_1, s_2, \ldots, s_n)$, we employ BERT to obtain contextualized token representations:

$$H_t = \mathrm{BERT}(S),$$

where $H_t \in \mathbb{R}^{n \times d_t}$ and d_t denotes the textual feature dimension.

To facilitate subsequent cross-modal interactions, the textual and visual representations are projected into a shared feature space. Specifically, we apply linear projections to the textual and visual representations, respectively:

$$H_t^{'} = \mathrm{Project}_t(H_t)$$

$$h_v^{'} = \mathrm{Project}_v(H_v)$$

where $H_t^{'} \in \mathbb{R}^{n \times d}$, $h_v^{'} \in \mathbb{R}^{d}$, and d denotes the shared hidden dimension.

Since the visual representation is a sample-level vector whereas the textual representation is a token-level matrix, the visual vector is repeated along the token dimension to obtain a visual representation matrix aligned with the input length:

$$H_v^{'} = \mathrm{Repat}(h_v^{'}, n)$$

where $H_v^{'} \in \mathbb{R}^{n \times d}$.

After these operations, the textual and visual modalities are mapped into the same feature space, which enables subsequent consistency estimation and progressive regulation.

E. Visual evidence reliability estimator

To explicitly characterize how strongly the visual modality supports the current sample, we design a visual evidence reliability estimator. Here, “reliability” does not refer to the intrinsic quality of the image itself, but rather to the task relevant support provided by visual information for the current textual semantics and entity recognition task.

$$H_t^{'} = [h_{t,1}^{'}, h_{t,2}^{'}, \ldots, h_{t,n}^{'}],$$

where $h_{t,i}^{'} \in \mathbb{R}^{d}$ denotes the textual representation of the i-th token, and $h_v^{'} \in \mathbb{R}^{d}$ denotes the sample-level visual representation. We first compute the similarity between each token representation and the visual representation:

$$e_i = h_{t,i}^{'\top} h_v^{'}, \quad i \in \{1,2,\ldots,n\}$$

The similarities over all tokens are then averaged, and a nonlinear function is applied to suppress the effect of negative values:

$$r = \mathrm{ReLU}\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right)$$

Finally, the visual evidence reliability score is obtained through a sigmoid function:

$$\alpha = \sigma(r)$$

where $\alpha \in [0,1]$.

A larger α indicates that the visual information is more consistent with the textual semantics of the current sample and is therefore more supportive for entity recognition. In contrast, a smaller α suggests that the contribution of visual information to the current task is limited, and its influence should be reduced during subsequent fusion. By explicitly introducing this score, the model can adaptively regulate the contribution of the visual modality across different samples.

F. Progressive reliability regulation module

After obtaining the visual evidence reliability score, we further design a progressive reliability regulation module to progressively calibrate cross-modal interactions. This module is composed of multiple stacked progressive reliability regulation units, which continuously regulate the influence of the visual branch on multimodal representation updating across layers.

1) Initial Multimodal State: We first construct the initial multimodal state by combining the textual and visual representations:

$$H_c^0 = H_t^{'} + H_v^{'}$$

where $H_c^0 \in \mathbb{R}^{n\times d}$. This representation serves as the input to the subsequent progressive regulation process.

2) Progressive Reliability Regulation Unit: Let H_c^{l-1} denote the multimodal state output from the (l-1)-th layer. In the l-th regulation unit, the inputs include H_c^{l-1}, the textual representation $H_t^{'}$, the visual representation $H_v^{'}$, and the visual evidence reliability score α.

We first compute the cross-modal updates induced by the textual branch and the visual branch on the current multimodal state:

$$\widetilde{H}_{t\to c}^{l} = \mathrm{LN}\left(H_c^{l-1} + \mathrm{CAttn}\left(Q(H_c^{l-1}), K(H_t^{'}), V(H_t^{'})\right)\right),$$
$$\widetilde{H}_{v\to c}^{l} = \mathrm{LN}\left(H_c^{l-1} + \mathrm{CAttn}(Q(H_c^{l-1}), K(H_v^{'}), V(H_v^{'}))\right).$$

where $\mathrm{CAttn}(\cdot)$ denotes the cross-modal attention mechanism and $\mathrm{LN}(\cdot)$ denotes layer normalization.

On this basis, the visual branch is further regulated by the visual evidence reliability score:

$$H_{v\to c}^{l} = \mathrm{LN}\left(\widetilde{H}_{v\to c}^{l} + \alpha \cdot \widetilde{H}_{v\to c}^{l}\right),$$
$$H_{t\to c}^{l} = \mathrm{LN}\left(\widetilde{H}_{t\to c}^{l}\right).$$

The two branches are then fused to produce the calibrated output of the current layer:

$$H_f^{l} = H_c^{l-1} + \mathrm{LN}\left(H_{t\to c}^{l} + H_{v\to c}^{l}\right)$$

Finally, the fused representation is passed through a multi-head self-attention layer and a feed-forward network to obtain the multimodal state of the next layer:

$$H_c^{l} = \mathrm{FFN}\left(\mathrm{MHA}\left(H_f^{l}\right)\right)$$

where MHA(·) denotes multi-head self-attention and FFN(·) denotes the feed-forward network.

By stacking L progressive reliability regulation units, the model ultimately obtains the deep multimodal representation:

$$H_c^{L} = \mathrm{PRRM}(H_t^{'}, H_v^{'}, \alpha)$$

where PRRM(·) denotes the progressive reliability regulation module.

3)Module Characteristics: The advantages of the progressive reliability regulation module are mainly reflected in two aspects. First, the model separately addresses “whether the visual evidence is reliable” and “how visual evidence should be incorporated into fusion,” with these two roles handled by the visual evidence reliability estimator and the progressive reliability regulation module, respectively. This makes the fusion process more explicit and interpretable. Second, instead of injecting visual information only once, the model

continuously calibrates the contribution of the visual branch throughout multi-layer cross-modal interactions, allowing the role of the visual modality to evolve dynamically with the multimodal representation.

G. CRF decoding and training objective

After the progressive reliability regulation module, the final multimodal representation H_c^L is obtained. Based on this representation, a CRF decoder is employed for sequence labeling:

$$p(Y|H_c^L) = \mathrm{CRF}(H_c^L)$$

The sequence scoring function is defined as:

$$\mathrm{score}(H_c^L, Y) = \sum_{i=1}^{n} E_{i,y_i} + \sum_{i=0}^{n} T_{y_i,y_{i+1}}$$

where E_{i,y_i} denotes the emission score of assigning label y_i to the i-th token, and $T_{y_i,y_{i+1}}$ denotes the transition score between adjacent labels.

The model is trained by minimizing the negative conditional log-likelihood:

$$\mathcal{L} = -\log p(Y|H_c^L)$$

By introducing the CRF layer, the model can further exploit transition dependencies among labels, thereby improving the consistency and accuracy of sequence labeling.

IV. Experiments

A. Datasets and evaluation metrics

We conduct experiments on two publicly available Multi-image multimodal named entity recognition datasets, namely MNER-MI and MNER-MI-Plus. MNER-MI is a benchmark specifically constructed for Multi-image MNER, where each sample typically contains 2–4 images associated with the input text. Building upon this setting, MNER-MI-Plus further incorporates both single-image and Multi-image samples, covering a wider range of image–text combinations. Both datasets follow standard training, validation, and test splits, and the entity categories include Person (PER), Location (LOC), Organization (ORG), and Miscellaneous (MISC).

For evaluation, we adopt Precision (P), Recall (R), and F1-score, with F1-score serving as the primary metric for comparison.

B. Experimental settings

For text encoding, we use BERT-base, while the visual encoder is ViT-base-patch16. Given Multi-image input, the model first encodes each image independently and then derives a joint visual representation through the inter-image interaction module. AdamW is used as the optimizer. The learning rate is searched over {1e-5,2e-5,3e-5,4e-5, 5e-5,7e-5}, and the batch size is selected from {8,16,32}. The best hyperparameter configuration is selected on the validation set, and the final results are reported on the test set.

C. Baselines

To comprehensively evaluate the effectiveness of the proposed method, we compare it against text-only methods, single-image multimodal methods, and Multi-image multimodal methods.

Text-only methods include BiLSTM-CRF, CNN-BiLSTM-CRF, HBiLSTM-CRF, BERT, and GPT4. These methods rely solely on textual information for entity recognition and are used to measure the gains brought by visual information.

Single-image multimodal methods include GVATT-HBiLSTM-CRF, AdaCAN-CNN-BiLSTM-CRF, UMT, MAF, UMGF, VisualPT-MoE, HVPNeT, and MiniGPT4. These approaches are primarily designed for single-image settings. In our experiments, multiple images are first unified into a single visual input format before being fed into these models.

Multi-image multimodal methods include UMT-MI, UMGF-MI, VisualPT-MoE-MI, TPM-MI,

GIFR, and MI-Grap. These methods can directly handle Multi-image input and therefore serve as the primary baselines for comparison.

D. Main results

Table 1 reports the performance of the proposed method and all baselines on MNER-MI and MNER-MI-Plus.

Overall, text-only methods can model linguistic context effectively, but their performance remains inherently limited in Multi-image scenarios due to the absence of complementary visual information. Among these methods, BERT achieves the best performance, indicating the strong capability of pretrained language models in social media named entity recognition.

Table 1 Performance comparison of different models on mner-mi and mner-mi-plus datasets

Modality	Methods	Twitter 15			Twitter 17		
		Pre	Rec	F1	Pre	Rec	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 + Single Image	MiniGPT4	59.87	62.37	61.09	62.22	64.27	63.23
	GVATT-HBiLSTM-CRFA	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
	HPVNeT	74.93	75.28	75.10	81.88	80.94	81.41
Text + Multiple Images	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
	VERC-MI	75.53	81.19	78.22	81.80	83.72	82.75

Compared with text-only baselines, single-image multimodal methods generally achieve better results, which confirms that visual information can indeed provide useful complementary cues for entity recognition. However, because these methods are mainly designed for single-image settings, they often struggle to fully handle inter-image variation and redundancy when multiple images are available, leading to only limited improvements in Multi-image scenarios.

Multi-image multimodal methods further exploit the complementary semantics provided by multiple images and generally outperform single-image methods. In particular, TPM-MI, GIFR, and MI-Grap all achieve strong performance, suggesting that explicitly modeling inter-image relations is beneficial for improving MNER in Multi-image settings.

The proposed method achieves competitive results on both datasets. On MNER-MI, it obtains 77.74, 80.86, and 79.27 in Precision, Recall, and F1-score, respectively. On MNER-MI-Plus, it reaches 82.55, 84.47, and 83.50, respectively. These results demonstrate that jointly encoding multiple images and introducing visual evidence reliability modeling during fusion can effectively enhance performance on Multi-image MNER. Compared with static fusion strategies, the proposed method dynamically adjusts the contribution of visual information according to image–text consistency, thereby making better use of complementary visual cues across multiple images.

E. Ablation study

To evaluate the contribution of each component, we conduct ablation experiments, and the results are reported in Table 2. Specifically, we consider the following three variants:

• w/o progressive reliability regulation module: the progressive reliability regulation module is removed and replaced with a simpler static fusion strategy;

• w/o CRF: the CRF decoder is replaced with a standard softmax classifier;

• Full model: the visual evidence reliability estimator, the progressive reliability regulation module, and the CRF decoder are all retained.

As shown in Table 2, removing the progressive reliability regulation module causes the F1-score to drop to 78.26 and 82.75 on the two datasets, respectively. This indicates that simple fusion alone is insufficient to capture the varying importance of visual evidence across samples. The result confirms that the progressive reliability regulation module can continuously adjust the contribution of the visual branch throughout multi-layer interactions, thereby improving multimodal fusion.

Replacing the CRF decoder with a softmax classifier also leads to performance degradation, suggesting that label transition dependencies remain beneficial for final entity recognition.

Overall, the full model achieves the best results on both datasets, demonstrating that both visual evidence reliability modeling and sequence-level label modeling are essential for Multi-image MNER.

Table 2 Ablation study of verc-mi on the mner-mi and mner-mi-plus datasets

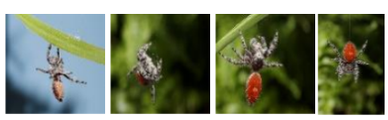
Model	MNER-MI			MNER-MI-Plus		
	P	R	F1	P	R	F1
VERC-MI	75.53	81.19	78.26	81.80	83.72	82.75
w/o PRRM	72.28	79.67	75.88	79.14	78.31	79.90
w/o-CRF	75.15	78.91	77.11	80.27	78.86	81.57

F. Case study

To further illustrate the effectiveness of the proposed method, Table 3 presents two representative examples. In the first case, the text contains the pronoun her, which may interfere with entity type prediction and mislead the baseline TPM-MI into classifying *Phidippus adumbratus* as **PER**. However, the accompanying image clearly show that the entity refers to a kind of spider. This suggests that, in this sample, visual information is more reliable than textual clues. Our method is able to make better use of the visual evidence and therefore correctly identifies the entity as **MIS**. In the second case, the entity Auto Select appears in an advertising context, where the text itself already provides relatively clear semantic cues. Therefore,

in this sample, textual information plays the dominant role, while visual information serves as auxiliary support. Our method correctly identifies it as Organization, whereas TPM-MI incorrectly predicts **O**. These two examples show that, in multi-image named entity recognition, different samples rely on textual and visual modalities to different degrees. The proposed method can dynamically adjust the contribution of each modality according to the characteristics of the sample, thereby improving entity recognition accuracy.

Table 3 Case study of two representative examples

No	Text and Images	Gold	TPM-MI	CODF
1	Phidippus adumbratus is a jumpy little jumper, here she is hanging out. 	Phidippus adumbratus (MIS)	Phidippus adumbratus (PER) ✗	Phidippus adumbratus (MIS) ✓
2	Check out may specials at auto select! click here for more info.	Auto Select (Organization)	Auto Select (O) ✗	Auto Select (Organization) ✓

V. Conclusion

The above results suggest that the effectiveness of the proposed method mainly comes from two aspects. On the one hand, Multi-image visual encoding together with inter-image interaction enables the model to integrate information from multiple images at the sample level, thus providing more complete visual support for entity recognition. On the other hand, the visual evidence reliability estimator and the progressive reliability-aware regulation mechanism allow the model to dynamically adjust the contribution of the visual modality according to image–text consistency, thereby making multimodal fusion more targeted and effective.

These findings indicate that, for Multi-image MNER, simply introducing more images is not sufficient to guarantee performance gains. The key lies in how to effectively exploit these visual inputs in a sample-adaptive manner. By explicitly modeling visual evidence reliability, the proposed method offers a more direct solution for image–text fusion in Multi-image scenarios.

Acknowledgment

This work is supported by Yunnan Provincial Science and Technology Program Project (202501BA070001-109), Yuxiu Innovation Project of NCUT (2024NCUTYXCX102), Youth Research Special Project of NCUT (2025NCUTYRSP012)).

References

[1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//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). 2019: 4171-4186.

[2] Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-

level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.

[3] Xu D, Chen W, Peng W, et al. Large language models for generative information extraction: A survey[J]. Frontiers of Computer Science, 2024, 18(6): 186357.

[4] Zhang Q, Fu J, Liu X, et al. Adaptive co-attention network for named entity recognition in tweets[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[5] Yu J, Jiang J, Yang L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3342-3352.

[6] Lu D, Neves L, Carvalho V, et al. Visual attention model for name tagging in multimodal social media[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1990-1999.

[7] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1715-1725.

[8] Chen D, Li Z, Gu B, et al. Multimodal named entity recognition with image attributes and image knowledge[C]//International Conference on Database Systems for Advanced Applications. 2021: 186-201.

[9] Zhang D, Wei S, Li S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(16): 14347-14355.

[10] Wu Z, Zheng C, Cai Y, et al. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1038-1046.

[11] Huang S, Xu B, Li C, et al. MNER-MI: A multi-image dataset for multimodal named entity recognition in social media[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 11452-11462.

[12] Qian S, Jin W, Chen Y, et al. A survey on multimodal named entity recognition[C]//International Conference on Intelligent Computing. 2023: 609-622.

[13] Zeng Q, Yuan M, Su Y, et al. Improving multimodal named entity recognition via text-image relevance prediction with large language models[J]. Neurocomputing, 2025: 130982.

[14] Mu Y, Guo Z, Li X, et al. MCIRP: A multi-granularity cross-modal interaction model based on relational propagation for multimodal named entity recognition with multiple images[J]. Information Processing & Management, 2026, 63(2): 104384.

[15] Zhao J, Huang S, Lin X. A graph interaction framework on relevance for multimodal named entity recognition with multiple images[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 1237-1246.

[16] Xu B, Huang S, Sha C, et al. MAF: A general matching and alignment framework for multimodal named entity recognition[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1215-1223.

[17] Xu B, Huang S, Du M, et al. Different data, different modalities! Reinforced data splitting for effective multimodal information extraction

from social media posts[C]//Proceedings of the 29th International Conference on Computational Linguistics. 2022: 1855-1864.

[18] Wang G, Liu J, Xie J, et al. Joint multimodal entity relation extraction based on temporal enhancement and similarity-gated attention[J]. Knowledge-Based Systems, 2024, 304: 112504.

[19] Bao X, Tian M, Zha Z, et al. MPMRC-MNER: A unified MRC framework for multimodal named entity recognition based multimodal prompt[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023: 47-56.

[20] Yu J, Li Z, Wang J, et al. Grounded multimodal named entity recognition on social media[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 9141-9154.

[21] Mai W, Zhang Z, Li K, et al. Dynamic graph construction framework for multimodal named entity recognition in social media[J]. IEEE Transactions on Computational Social Systems, 2023, 11(2): 2513-2522.

[22] Hu J, Lyu Y, Xue Y. VS-MRC: A visual semantics-guided machine reading comprehension framework for multimodal named entity recognition with multiple images[J]. Knowledge-Based Systems, 2025, 326: 114024.

[23] Chen X, Zhang N, Li L, et al. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction[C]//Findings of the Association for Computational Linguistics: NAACL 2022. 2022: 1607-1618.

[24] Wu J, Gong C, Cao Z, et al. MCG-MNER: A multi-granularity cross-modality generative framework for multimodal NER with instruction[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 3209-3218.

[25] Xu B, Huang S, Du M, et al. A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction[C]//International Conference on Database Systems for Advanced Applications. 2023: 544-554.

[26] Sun L, Wang J, Zhang K, et al. RPBERT: A text-image relation propagation-based BERT model for multimodal NER[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13860-13868.

[27] Jia M, Shen L, Shen X, et al. MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(7): 8032-8040.

[28] Wu Y, Liu J. Multimodal named-entity recognition based on symmetric fusion with contrastive learning[J]. Symmetry, 2026, 18(2): 353.

[29] Chen F, Feng Y. Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction[EB/OL]. 2023.

[30] Zha E, Zeng D, Lin M, et al. CEPTNER: Contrastive learning enhanced prototypical network for two-stage few-shot named entity recognition[J]. Knowledge-Based Systems, 2024, 295: 111730.

[31] Yang K, Yang Z, Zhao S, et al. Uncertainty-aware contrastive learning for semi-supervised named entity recognition[J]. Knowledge-Based Systems, 2024, 296: 111762.

[32] Zhu D, Chen J, Shen X, et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models[EB/OL]. 2023.

©2026. This article is copyrighted by the author and Hong Kong Science and Technology Publishing Group. This work is licensed under a Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/

Open Access