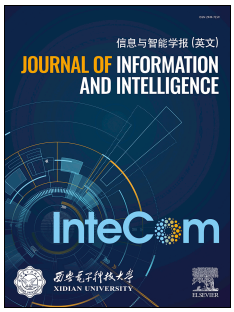
Journal Pre-proof



Deep Modular Skip-Attention Networks for Multimodal Named Entity Recognition

Chengen Lai, Shengli Song, Sitong Yan, Guangneng Hu

PII: S2949-7159(26)00035-1

DOI: https://doi.org/10.1016/j.jiixd.2026.04.004

Reference: JIIXD 140

To appear in: *Journal of Information and Intelligence*

Received Date: 31 August 2025

Revised Date: 29 January 2026

Accepted Date: 15 April 2026

Please cite this article as: Lai C., Song S., Yan S. & Hu G., Deep Modular Skip-Attention Networks for Multimodal Named Entity Recognition, *Journal of Information and Intelligence*, https://doi.org/10.1016/j.jiixd.2026.04.004.

Deep Modular Skip-Attention Networks for Multimodal Named Entity Recognition

Chengen Lai[a], Shengli Song[a,*], Sitong Yan[a], Guangneng Hu[a]

[a]*School of Computer Science and Technology, Xidian University, Xi'an, 710126, Shaanxi, China*

Abstract

Multimodal named entity recognition (MNER) has achieved significant progress in recent years. However, existing approaches still suffer from two drawbacks. Firstly, for interaction effectiveness, the shallow multimodal interaction models fail in capturing fine-grained interactions between multimodal instances and suffer from limited modality information. The deep models that use simple concatenation strategies demand a huge amount of computation resources while achieving better performance. Secondly, for interaction strategies, if the associated image is noisy and unrelated to the text, multimodal representation in MNER is often biased and even omits the important information from the dominant modality in current fusion methods, which makes them suffer from visual bias.

To address the above issues, we proposed DESER (**DE**ep Modular **S**kip-Attention Networks for MN**ER**) to effectively model the interactions across modalities and debias the noisy modality impact. Specifically, the novel Skip-Attention layer effectively models the interactions across modalities and preserves the important information with text guidance by cascading in depth while reducing resource consumption. The multimodal gating fusion module adaptively distinguishes dominant modalities from all modalities and controls the contributions of the unrelated modalities by utilizing self-optimizing pseudo-label training to debias the noisy modality impact. Empirical results on two benchmark datasets show the effectiveness of DESER.

*Corresponding author

Email addresses: laice@stu.xidian.edu.cn (Chengen Lai), shlsong@xidian.edu.cn (Shengli Song), styan@stu.xidian.edu.cn (Sitong Yan), njuhgn@gmail.com (Guangneng Hu)

Deep Modular Skip-Attention Networks for Multimodal Named Entity Recognition

Chengen Lai[a], Shengli Song[a,*], Sitong Yan[a], Guangneng Hu[a]

[a]*School of Computer Science and Technology, Xidian University, Xi'an, 710126, Shaanxi, China*

Abstract

Multimodal named entity recognition(MNER) has achieved significant progress in recent years. However, existing approaches still suffer from two drawbacks. Firstly, for interaction effectiveness, the shallow multimodal interaction models fail in capturing fine-grained interactions between multimodal instances and suffer from limited modality information. The deep models that use simple concatenation strategies demand a huge amount of computation resources while achieving better performance. Secondly, for interaction strategies, if the associated image is noisy and unrelated to the text, multimodal representation in MNER is often biased and even omits the important information from the dominant modality in current fusion methods, which makes them suffer from visual bias.

To address the above issues, we proposed DESER (**DE**ep Modular **S**kip Attention Networks for MN**ER**) to effectively model the interactions across modalities and debias the noisy modality impact. Unlike shallow cross-modal attention models that capture only single-layer interactions, and deep concatenation-based frameworks that indiscriminately fuse modalities at all layers, DESER introduces a modular skip-attention architecture that enables deep, text-guided multimodal interaction while explicitly preserving dominant textual semantics. The proposed Skip-Attention layer creates inter-layer shortcuts that selectively bypass visual adaptation, allowing the model to balance interaction depth and computational efficiency. Furthermore, dif-

*Corresponding author

Email addresses: laice@stu.xidian.edu.cn (Chengen Lai), shlsong@xidian.edu.cn (Shengli Song), styan@stu.xidian.edu.cn (Sitong Yan), njuhgn@gmail.com (Guangneng Hu)

(a) [**SheilaS PER**] finally made it to a special concert, the Symphony in [**the Flint Hills LOC**].

(b) Me trying to explain [**Harry Potter MISC**] to a muggle

Figure 1: **Examples of multimodal named entity recognition (MNER). The named entities and their corresponding entity types are in bold. (a) A relevant text-image pair, (b) A hard, noisy pair.**

ferent from existing fusion methods that assume visual information is always beneficial, DESER incorporates a multimodal gating fusion module with self-optimizing pseudo-label training, which explicitly models modality dominance and suppresses noisy or irrelevant visual signals.

Keywords: Named Entity Recognition, Multimodal learning, Deep skip-attention network, Social media understanding

1. INTRODUCTION

Multimodal Named Entity Recognition (MNER) aims to identify and classify named entities from text accompanied by visual context. Unlike unimodal NER, MNER requires the model to reason over heterogeneous modalities with fundamentally different semantic structures and noise characteristics. For example, one can recognise "Rocky" as the name of a dog according to the accompanying image of the post. To incorporate visual information for recognizing named entities in multimodal posts, the core issue is to model the multimodal interaction across the text and image modalities. As shown in Figure 1(a), the image is segmented into four regions while the text is a separated sequence of tokens. To correctly identity the entity type of "SheilaS" is a named person, MNER must learn the multimodal interaction relationship between "SheilaS" and the visual features from upper right part of the image. Existing MNER methods make various attempts in achieving this goal.

Neural MNER approaches have been proposed to model mulitmodal interactions. Attention-based methods are firstly adopted to combine text and

image representations adaptively [1], including the LSTM-CNN architecture [2], visual attention and gating mechanism [3], and adaptive co-attention network [4]. Since Transformer and BERT-based methods have shown an effective performance on sequence learning [5, 6], they have experienced a rapid development in MNER. For example, multimodal transformers [7, 8] incorporated crossmodal attention to learn image-aware text representation and further the dual-side multimodal transformer [9] to learn word-aware visual representation as well. Deep neural MNER methods are recently proposed to learn deep and fine-grained representations of text and image modalities. The core idea is about concatenating text and image together and then feed it to deep-layer networks to learn the multimodal interactions [10, 11], with the options of preprocessing the image to extract its text semantics like object tags and image captions using off-the-shelf computer vision tools [12]. The fine-grained interactions can be captured using graph-based fusion which represents the input sentence and image by a unified multi-modal graph [13]. However, if the associated image is noisy to the text, for example, as shown in Figure 1(b), the visual information (a teacher-like person) is irrelevant with the named entity (Harry Potter) in the text, deep neural MNER may have negative effect. This makes them suffer from visual bias and fail in exploiting effective interaction strategies.

In summary, although the aforementioned MNER methods have shown performance improvement, they still suffer from two weaknesses from the perspectives of multimodal interaction effectiveness and multimodal interaction strategies. Firstly, for multimodal interaction effectiveness, the shallow multimodal interaction models fail in capturing fine-grained interactions between multimodal instances and suffer from limited modality information (a single layer crossmodal Transformer shown in Figure 2 (a)). The deep models that use simple concatenation strategies demand a huge amount of computation resources while achieving better performance (shown in Figure 2 (b)). Secondly, for multimodal interaction strategies, if the associated image is unrelated to the text, MNER may bring in negative effect, and hence the simple multimodal fusion strategy is not enough to debias the noisy visual impact. The text modality as dominant modality has not been well-exploited by existing methods and this makes them suffer from visual bias and error propagation from the noisy image modality during deep multimodal fusing.

In this paper, we proposed DESER (**DE**ep Modular **S**kip-Attention Networks for MN**ER**) to tackle the challenges of enabling effective multimodal interaction effectiveness and strategies together. The novel Skip-Attention

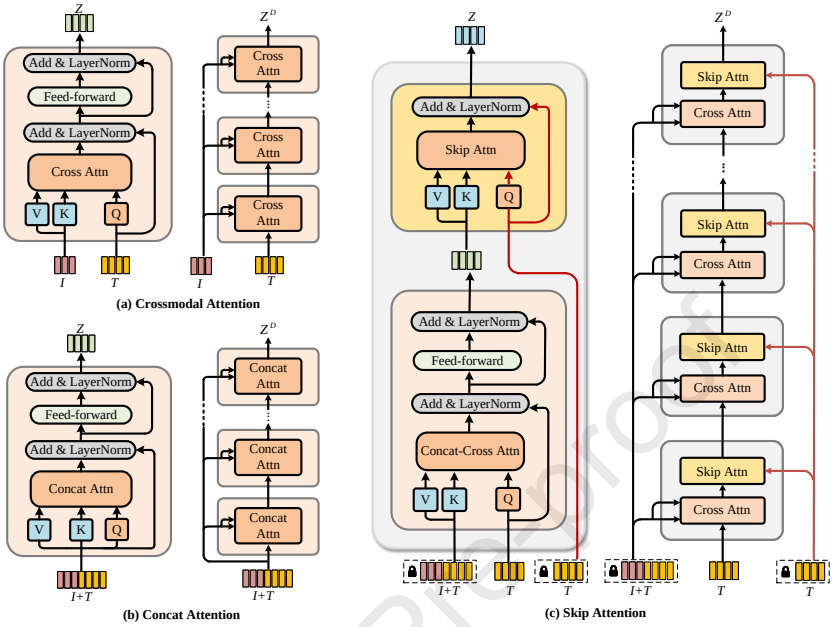


Figure 2: **Illustration of two existing multimodal fusion networks and the proposed skip attention network, where the lelf side of each subfigure is a single layer attention, and the right side is the corresponding multiple layers attention network. (a) Crossmodal Attention takes text and image as separated input and learns latent adaptation from one modality to the other. (b) Concat Attention adopts an early fusion mulitmodal interaction strategy which simply takes the concatenation of visual and text features as input to Transformer. (c) The Skip Attention creates inter-layer shortcuts skipping a crossmodal attention layer for visual adaptation to text modality which preserves the important textual information while reducing resource consumption.**

layer effectively models the interactions across modalities with text guidance. Instead of fusing visual and linguistic representations at the same levels, the skip attention creates inter-layer shortcuts by skipping a crossmodal attention layer for visual adaptation to text modality (shown in Figure 2 (c)). This achieves goals of both modeling the interactions across modalities and preserving the important textual information while reducing resource consumption. The multimodal gating fusion module adaptively distinguishes dominant modalities from all modalities and controls the contributions of the unrelated modalities by utilizing self-optimizing pseudo-label training. The differences between crossmodal attention, concat attention and our skip attention are summarized in Table1.

Table 1: Comparative Analysis of Multimodal Fusion Frameworks for MNER.

Multimodal Fusion Strategy	Fusion Depth	Modality Control	Design Objective
Crossmodal Attention	Shallow (1–2 layers)	Implicit	Inject visual cues into text
Concat Attention	Deep (full stacking)	None	Maximize multimodal interaction
Skip Attention	Deep (full stacking)	Explicit (gating + pseudo-labels)	Text-guided interaction with noise debiasing

Our main contributions can be summarized as follows:

- Wepropose DESER, adeep modular multimodal framework for MNER that enables text-guided multimodal interaction while preserving dominant textual semantics, addressing the trade-off between interaction depth and noise robustness.
- We introduce a Skip-Attention mechanism that allows selective visual information injection through inter-layer shortcuts, achieving deep multimodal fusion with reduced computational overhead.
- We design a multimodal gating fusion module supervised by self-optimizing pseudo-labels, which explicitly models modality dominance and mitigates the impact of noisy or irrelevant visual inputs.
- Extensive experiments on multiple benchmarks, demonstrate the effectiveness, robustness, and generalization ability of the proposed frame work.

2. RELATED WORK

We review related work on multimodal NER and skip-connections.

Named entity recognition Traditional named entity recognition (NER) methods design effective features and feed them into classifiers of maximum entropy, SVM, and CRF [14]. To reduce feature engineering cost, deep learning methods are proposed to couple with a CRF decoder layer, including CNN, RNN, Bidirectional NNs, and their hierarchical variants [15, 16, 17, 18]. Unsupervised, self-supervised, metric learning, and prompt-based learning are also adopted for NER task [19, 20]. These approaches achieve state-of-the-art performance on NER benchmark with formal, newswire text datasets. For social media user-generated text contents, however, their performance

gets much worse since social media posts are usually short, informal and linguistic variations.

Multimodal NER MNER methods are proposed to incorporate visual contexts where different neural architectures are adopted including LSTM-CNN, visual attention and co-attention, object-aware bilinear attention, prompt-based and pre-training networks [3, 21, 2, 22, 4, 23, 24]. Since Transformer and BERT have shown an effective performance on sequence learning, their shallow multimodal variants [25] have been proposed to incorporate cross-modal attention to learn image-aware text representation [7, 8, 9], and furthermore, few attempts to design deep architectures are recently explored [10]. External knowledge and extra image attributes are also incorporated into MNER [26, 27, 28]. However, they require external data sources [11], extensive computer vision tools to preprocess images [12, 29], extensive auxiliary task to enhance representations of text and image [30, 31], and tricky domain-specific graph construction [13, 32].

In this work, we propose the DESER to effectively model the interactions across modalities and debias the noisy modality impact. Specifically, the novel Skip-Attention layer effectively models the interactions across modalities and preserves the important information with text guidance by cascading in depth while reducing resource consumption. The multimodal gating fusion module adaptively distinguishes dominant modalities from all modalities and controls the contributions of the unrelated modalities by utilizing self-optimizing pseudo-label training to debias the noisy modality impact.

3. THE PROPOSED DESER MODEL

In this section, we give the overview architecture of the proposed DESER model and then introduce the details of each module. Firstly, we introduce the problem statement of MNER.

Problem Formulation Given a pair of sentence S and associated image V as input,the goal of MNER is to detect a set of entities from S and classify them into pre-defined types. As the same with existing MNER work, we formulate the task as a sequence labeling problem. Let $S = (s_1, s_2, \ldots, s_n)$ denote an n-sized sequence of input words, and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be the corresponding label sequence, where and is the pre-defined label set with *BIO2* tagging schema [33].

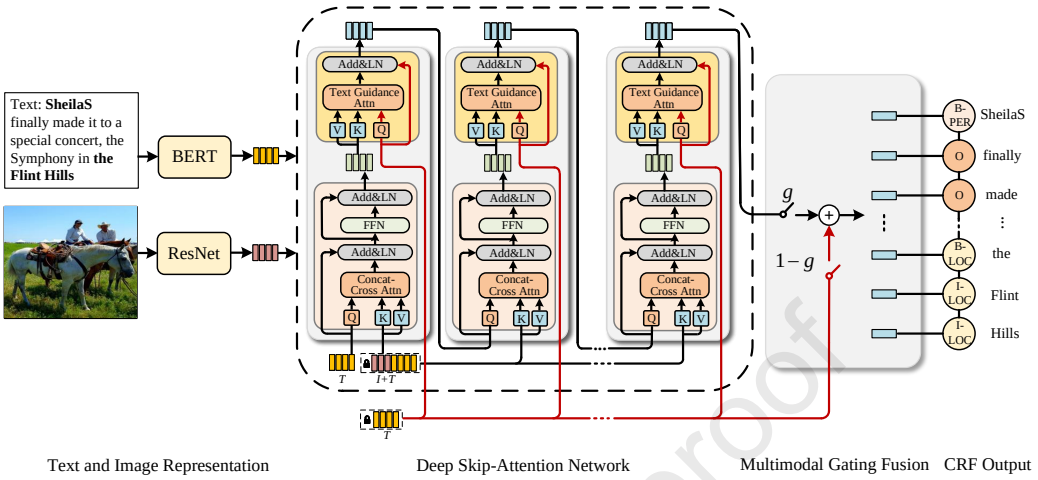


Figure 3: The overall architecture of the proposed DESER model. The flowchart consists of four parts. i) The Text and Image Representation module encodes text representation by BERT and image representation by ResNet. ii) The Deep Skip-Attention Network is to model multimodal interactions by stacking multiple Skip-Attention layers (see Figure 2), which can capture the fine-grained deep interactions between text and image modalities. iii) The Multimodal Gating Fusion module fuses text representation and multimodal representation by a gating mechanism with pseudo-label training. iv) The CRF Output module decodes the final fused hidden representation to predict entity types by a CRF layer.

3.1. Overall Architecture

The overall architecture of the proposed DESER model is shown in Figure 3. DESER consists of four modules. The first module is the text and image representation which obtains text representation by BERT model and image feature by ResNet model from the input text and associated image respectively. The second module is the deep skip-attention network which models deep multimodal interactions by following the text guidance leader through stacking the novel skip-attention layers in depth. The third and the fourth modules are the multimodal gating fusion and CRF decoder modules which adaptively fuse text feature and hidden representation by a gating network with pseudo-label training and then feed it to a CRF layer to predict entity types. We describe these modules in detail next.

3.2. Text and Image Representation

Text representation Due to the capability of learning different representations for the same word in different contexts, we use BERT [5] as the text

encoder. For a sequence of n-sized tokens $S = (s_1, \ldots, s_n)$, two special tokens are inserted into S, i.e.,appending [CLS] to the beginning and [SEP] to the end to obtain the BERT input $S' = (s_0, s_1, \ldots, s_{n+1})$, where s_0 and s_{n+1} denote the two special tokens respectively. Then we feed it into BERT to obtain the text representation $T = (t_0, t_1, \ldots, t_{n+1})$, where $t_i \in \mathbb{R}^d$ is the d-dimensional representation of token s_i

Image representation Due to the capability of extracting meaningful feature representations from image, we use ResNet [34] as the image encoder. Following [9] to obtain the spatial features of different regions, we split each input image into $7 \times 7 = 49$ visual blocks. An image V is firstly resized to 224×224 pixels and then fed into a pretrained ResNet to extract the image representation $\boldsymbol{V} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_{49}})$, where $\boldsymbol{v_i} \in \mathbb{R}^{2048}$ is the representation for the i-th visual blocks. To match the visual representation with the text representation, we convert V with a project layer to get the image representation $\boldsymbol{I} = \boldsymbol{W_I^\top V}$, where $\boldsymbol{W_I} \in \mathbb{R}^{2048 \times d}$ is the linear transformation matrix. The use of localized block-based representations rather than a global feature vector enables the model to attend to specific regions relevant to entity mentions and leverage visual attention mechanisms, where entity spans in the text can be conditioned on corresponding visual areas, which captures fine-grained visual-textual relationships. While this image representation approach is effective, fixed grid splitting may not perfectly align with semantically meaningful regions (e.g., faces, text signs), potentially introducing irrelevant visual noise. To address the challenge in image representation, we proposed the deep skip-attention network and multimodal gating fusion method.

3.3. Deep Skip-Attention Network

The deep skip-attention network consists of multiple skip attention layers. In skip attention layer, we first adopt concatenation strategy to form the input of key and value of a crossmodal attention layer and then create inter-layer shortcut by skipping the crossmodal attention layer for visual adaptation to text modality where the text is to form the input query of another multihead attention layer (see Figure 2). In this way, a skip attention layer preserves the important textual information and reduces resource consumption during deep multimodal fusion when it is cascaded in depth to model deep multimodal interactions. We will go into details of skip attention.

Multihead attention for sequence learning All of crossmodal attention, concat attention, and skip attention are based on multihead attention (see

Figure 2). The input of scaled dot-product attention in Transformer consists of queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{m\times d}$ and values $V \in \mathbb{R}^{m\times d}$ where n is the number of queries and m is the number of key-value pairs. The attended features $\boldsymbol{F}$ are calculated by weighted summation over values w.r.t. the attentions learned between queries and keys:

$$\boldsymbol{F} = Attn(Q, K, V) = softmax(\frac{QK^\top}{\sqrt{d}})V \tag{1}$$

To improve representation capacity of attended features, multihead attention consists of h paralleled "heads" where each head corresponds to an independent scaled dot-product attention:

$$\boldsymbol{F} = MultiAttn(Q, K, V) = [head_1, \ldots, head_h]W^O \tag{2}$$

$$head_i = Attn(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where $W_i^{Q,K,V} \in \mathbb{R}^{d\times d_h}$ are projection matrices for the i-th head, $W^O \in \mathbb{R}^{h*d_h\times d}$, and $d_h = d/h$ is the dimension of each head's output features. All projection matrices in the multi-head attention modules (including concat cross attention and skip-attention) are not shared across layers, allowing each skip-attention layer to learn layer-specific multimodal interaction patterns.

Cross attention for multimodal interaction The crossmodal attention takes text and image as separated input and learns latent adaptation from one modality to the other. The query, key and value are respectively calculated by:

$$\underline{CrossAttn} \quad Q = Z^{(\ell)}, K = \boldsymbol{I}, V = \boldsymbol{I} \tag{4}$$

where $Z^{(\ell)}$ is the output of the ℓ-th attention layer and $Z^0 = \boldsymbol{T}$. The computational complexity is $\mathcal{O}(n \cdot m \cdot d)$

Concat attention for multimodal interaction The concat attention adopts early fusion of mulitmodal interaction strategy which simply takes the concatenation of visual and text features as input to multihead attention. The query, key and value are respectively calculated by:

$$\underline{ConcatAttn} \quad Q = [\boldsymbol{I}; \boldsymbol{T}], K = [\boldsymbol{I}; \boldsymbol{T}], V = [\boldsymbol{I}; \boldsymbol{T}] \tag{5}$$

where $\boldsymbol{T}$ and $\boldsymbol{I}$ are representations of text and image (see Section 3.2), and $[\boldsymbol{I};\boldsymbol{T}]$ is the concatenation of image representation and text representation. The computational complexity is $\mathcal{O}((n+m)^2 \cdot d)$

Concat-Cross attention in skip-attention layer To save the huge computer computing resources in multimodal interaction, we propose the concat-cross attention in our skip-attention layer, it firstly adopts early fusion of mulitmodal interaction strategy like concat attention, and then learns latent adaptation from image modality to text modality like crossmodal attention. The query, key and value are respectively calculated by:

$$\underline{CatCrsAttn} \quad Q = Z^{(\ell)}, K = [\boldsymbol{I};\boldsymbol{T}], V = [\boldsymbol{I};\boldsymbol{T}] \tag{6}$$

where $\boldsymbol{T}$ and $\boldsymbol{I}$ are representations of text and image (see Section 3.2), and $K = V = [\boldsymbol{I};\boldsymbol{T}] \in \mathbb{R}^{(n+m)\times d}$ is the concatenation of image representation and text representation. Note that $Z^{(\ell)}$ is the output of the ℓ-th layer skip-attention network (see Eq. 10), and $Z^0 = \boldsymbol{T}$.

The output of the concat-cross attention in our skip-attention layer is achieved by feeding the output $\boldsymbol{F}$ of concat-cross multihead into two layer norms (LN) and feed-forward network (FFN):

$$Z_{CatCrs} = LN(FFN(LN(\boldsymbol{F}+\boldsymbol{T})) + LN(\boldsymbol{F}+\boldsymbol{T})) \tag{7}$$

where $\boldsymbol{F}$ is computed by Eq. 2 by feeding input from Eq. 6. The layer normalization is defined as following:

$$LN\left(x_i\right) = \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}, \mu_i = \frac{1}{d}\sum_{j=1}^{d} x_{i,j}, \sigma_i^2 = \frac{1}{d}\sum_{j=1}^{d}\left(x_{i,j} - \mu_i\right)^2 \tag{8}$$

Concat-cross attention models the inter-modal interaction between each token $\boldsymbol{t}_i \in \boldsymbol{T}$ in text and each region $\boldsymbol{v}_i \in \boldsymbol{I}$ in image, and intra-modal interaction between pairwise tokens $(\boldsymbol{t}_i, \boldsymbol{t}_j)$ in text. The concat-based interaction strategy allows crossmodal attention to learn multimodal interactions across modalities from the bottom level in an early fusion way. And the computational complexity is $\mathcal{O}((n+m) \cdot n \cdot d)$, which is significantly lower than the quadratic complexity over the concat attention when m is larger.

Skip attention for following text guidance leader To preserve the important textual information, we propose the skip attention for following text guidance leader. The skip-attention layer creates inter-layer shortcut by skipping the cross-modal attention layer (i.e., the concat-cross attention computed by Eq. 7) for visual adaptation to text modality where the text is to

form the input query of another multihead attention layer (see Figure 2). In way this, a skip attention layer follows the text guidance leader and preserves the important textual information during deep multimodal fusion.

As for the skip multihead attention in our skip-attention layer, the query, key and value are respectively calculated by:

$$\underline{SkipAttn} \quad Q = \boldsymbol{T}, K = Z_{CatCrs}, V = Z_{CatCrs} \tag{9}$$

where Z_{CatCrs} is computed by Eq. 7. Note that, these skip multihead projection matrices are different from those concat-cross multihead projection matrices as shown in Eq. 6.

The output of the skip-attention layer is achieved by feeding the output $\boldsymbol{F}$ of the skip multihead attention into one normalization layer:

$$Z_{skip} = LN(\boldsymbol{F} + \boldsymbol{T}) \tag{10}$$

where $\boldsymbol{F}$ is computed by Eq. 2 by feeding input from Eq. 9.

Stacking skip attention layer We stack D skip-attention layers in depth to model deep multimodal interactions. We denote by Z^D the final output of the deep skip-attention network module:

$$Z^D = Z^{(D)}_{skip} \tag{11}$$

3.4. Multimodal Gating Fusion

After obtaining the multimodal representation $\boldsymbol{Z^D}$ that preserves the important textual information, to learn the contributions of different modalities in multimodal fusion, we propose the Multimodal Gating Fusion module to control the contributions of noisy modality and prevent error propagation during multimodal fusion. Specifically, we first calculate the contribution weight between modalities of each text-image input to the MNER task prediction adaptively by gating network. And then we weighted sum of the all modalities, including text modality $\boldsymbol{T}$ and aligned multi-modality $\boldsymbol{Z^D}$ to obtain the multimodal representation $\boldsymbol{H}$ that reduces the omitting problem of important information from the dominant modality by assigning low contribution scores to the unrelated modality and high contribution scores to the dominant modality.

$$\boldsymbol{s} = Softmax([\boldsymbol{T}; \boldsymbol{Z^D}]W^s) \tag{12}$$

$$\boldsymbol{H} = \boldsymbol{s}^t \odot \boldsymbol{T} + \boldsymbol{s}^{\boldsymbol{z}} \odot \boldsymbol{Z}^D \tag{13}$$

where $W^s \in \mathbb{R}^{2d\times 2}$ are learned matrices. $s = [s^t, s^z] \in \mathbb{R}^{n\times 2}$ is the predicted contribution scores of different modalities on each input.

To improve the reliability of predicted contribution scores in multimodal gating fusion, besides implicitly capturing the contribution relations between different modalities, inspired by the priori knowledge that the quality of representation is closely related to the loss [35], i.e., loss is smaller and the ability of representation is better in the neural network, we design a pseudo-label training task by self-optimizing combination reweights between modalities. To avoid instability caused by unreliable loss estimates during early training epochs, pseudo-label generation is not applied from the beginning of training. Instead, DESER is first trained on a subset of the training data to obtain a reasonably converged model. Pseudo-labels are then generated using the remaining data based on the stabilized loss signals.This two-stage training strategy prevents noisy or uninformative loss values in early epochs from adversely affecting the gating supervision, thereby ensuring stable optimization. Specifically, we first use three-quarters of the training dataset to train the initial DESER model. Secondly, for the remaining quarter of the training data, we feed the text modality T and multimodal data Z^D that derived from initial DESER into the classifier to yield the $loss_i^t$ and $loss_i^z$ between label and ground-truth on each token i by cross-entropy, respectively. Thirdly, we take their losses into the pseudo label tagging module to obtain the pseudo contribution labels l_i^t, l_i^z of different modalities on each token i with Eq 14. By normalizing each modality loss with respect to their sum, the proposed formulation transforms absolute losses into relative contribution scores, which allows the gating module to learn a competitive and interpretable weighting between modalities. This normalization ensures that the modality with a lower loss receives a higher pseudo-label weight, while the less reliable modality is correspondingly suppressed. This design is inspired by the observation that, under identical supervision, a modality yielding a lower prediction loss provides a more discriminative representation for the current sample. Therefore, the normalized inverse-loss formulation serves as a soft reliability estimator rather than a hard supervision signal. Finally, the auxiliary loss of contribution score training is calculated by the Eq 15 between pseudo contribution scores and predicted pseudo contribution labels with mean squared error as loss function. The contribution score distinguishes the dominant modalities from different modalities by guiding the

training with pseudo labels.

$$l_i^t = 1 - \frac{loss_i^t}{loss_i^t + loss_i^z}; l_i^z = 1 - \frac{loss_i^z}{loss_i^t + loss_i^z} \tag{14}$$

$$\mathscr{L}_{pseudo} = \frac{1}{n}(\Sigma_{i=1}^{n}(s_i^t - l_i^t)^2 + \Sigma_{i=1}^{n}(s_i^z - l_i^z)^2) \tag{15}$$

3.5. CRF Output

After the final hidden representation $\boldsymbol{H}$ is obtained from the Multimodal Fusion module, we use conditional random field (CRF) decoder to perform the MNER task since CRF can produce higher tagging accuracy in sequence labelling learning by exploiting correlations between neighbouring labels. For example, I-PER cannot follow B-LOC in NER while B-LOC is often followed by I-LOC. Therefore, instead of decoding each label independently, we feed hidden representation $\boldsymbol{H}$ into a standard CRF layer to perform conditional sequence labelling $\boldsymbol{y}$ through the text sentence S and its associated visual image V as follow:

$$P(\boldsymbol{y} \mid S, V) = \frac{exp(score(\boldsymbol{H}, \boldsymbol{y}))}{\sum_{\boldsymbol{y'}} exp((score(\boldsymbol{H}, \boldsymbol{y'}))} \tag{16}$$

$$score(\boldsymbol{H}, \boldsymbol{y}) = \sum_{i=0}^{n} T_{y_i, y_{i+1}} + \sum_{i=1}^{n} E_{\boldsymbol{h}_i, y_i} \tag{17}$$

where $E_{\boldsymbol{h}_i, y_i}$ is the emission score of label y_i for the i-th token, and $T_{y_i, y_{i+1}}$ is the transition score from label y_i to label y_{i+1}.

Training To train the module, we use the negative log-likelihood as the loss function, which is defined as follows:

$$\mathscr{L}_{MNER} = -logP(\boldsymbol{y} \mid S, V) \tag{18}$$

$$\mathscr{L}_{Total} = \mathscr{L}_{MNER} + \mathscr{L}_{pseudo} \tag{19}$$

where $\mathscr{L}_{MNER}$ is the loss to train the initial DESER model on three-quarters of the training dataset. $\mathscr{L}_{Total}$ is the loss to train the DESER model with pseudo labels on remaining quarter of the training dataset.

Inference In the decoding phrase, we predict the output labeling sequence that achieves the maximum conditional score given by:

$$\boldsymbol{y}^* = argmax_{\boldsymbol{y'}} P(\boldsymbol{y'} \mid S, V) \tag{20}$$

Table 2: Hyperparameters of DESER when comparing with state-of-the-art (text-only/ multimodal) named entity recognition methods.

Hyperparameter	Setting
sentence encoder	BERT-base-uncase
image encoder	ResNet152
dimension of image feature	2048
dimension of text feature	768
maximum length of sentence	128
depth of skip-attention layers	12
batch size	8
total iterations	30
random seed	1236
activation	sigmoid
optimizer	AdamW
warmup ratio of optimizer	0.01
learning rate	3e-5
dropout rate	0.5

4. EXPERIMENTS

We conduct experiments on two benchmark datasets to extensively evaluate DESER by answering the following research questions:
RQ1: How does DESER perform by comparing with state-of-the-art methods for MNER?
RQ2: How does DESER perform on interaction effectiveness and interaction strategies?
RQ3: Are qualitative results helping understand how the proposed model works?

4.1. Data and Setup

Datasets We conduct empirical experiments on two widely used benchmark datasets, namely Twitter-15 [4] and Twitter-17 [3]. Each Tweet post contains a pair of (text, image), where text and image may be unrelated, and text can have zero named entity. The four types of entities and dataset statistics are shown in Table 3.
Metrics To evaluate the performance of MNER methods, we report the widely used three metrics: Precision, Recall, and F1 score over overall performance, and F1 score for each of the four types.
Parameter setting As shown in Table 2, we conduct all experiments on NVIDIA GTX 3080 Ti GPUs with PyTorch 1.9.0. BERT-base with hidden

size of 768 is used to encode text sentence while ResNet152 with dimension of 2048 is used to extract features from image. The depth of stacking is 12. The maximum length of text sentence cuts at 128, batch size is 8, and training epoch is 30. AdamW is the optimizer, learning rate and dropout rate for training are 3e-5 and 0.5, respectively.

4.2. Baselines

To demonstrate the effectiveness of the proposed model, we compare DESER with four text-only NER approaches, four shallow multimodal methods, and five deep MNER models.

The first group is four text-only NER approaches: **CNN-BiLSTM-CRF** [36] exploits character-level word representations. **HBiLSTM-CRF** [37] is a bidirectional LSTM followed by a CRF layer. **BERT** [5] is a multilayer bidirectional Transformer. **BERT-CRF** is a BERT followed by a CRF decoder.

The second group is four shallow multimodal methods: **VG** [3] is based on HBiLSTM-CRF and adopts a visual attention and gate to obtain multimodal fusion. **ACoA** [4] is based on CNN-BiLSTM-CRF and adopts an adaptive co-attention network to model textual and visual interactions. **UMT** [9] adopts crossmodal attention to model multimodal interaction with an auxiliary text-only learning task. **MAF** [8] uses a crossmodal matching module to control the proportion of image features and a crossmodal alignment module to learn unified representation of two modalities.

The third group is five deep MNER models: **RpBERT** [11] adopts concat attention to model multimodal interaction in an early fusion way. **UMGF** [13] uses a graph-based fusion to capture semantic relationships between words and visual objects. **HVPNeT** [10] uses a visual prefix to concatenate with the text representation. **ITA** [12] concatenates text with visual contexts (including object tags, image captions and optical characters) and then feed it into BERT. **ICKA** [28] leverages the knowledge from VLM to extract implicit interactions within text–image pairs.

4.3. Performance Comparison: RQ1

Main results The performance comparison of different methods is shown in Table 4 where the best results are in boldface and the second best underlined. We have following observations.

Table 3: Statistics of the two benchmark datasets.

Entity Type	Twitter-15			Twitter-17		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Person (PER)	2217	552	1816	2943	626	621
Location (LOC)	2091	522	1697	731	173	178
Organization (ORG)	928	247	839	1674	375	395
Miscellaneous (MISC)	940	225	726	701	150	157
Total Entity	6176	1546	5078	6049	1324	1351
Num of Tweets	4000	1000	3257	3373	723	723

Firstly, approaches exploit visual contexts get better performance than text-only NER methods. For example, the performance of shallow multimodal methods (UMT and MAF) is better than all text-only NER methods with a large margin improvement in terms of F1 metric on both datasets, pushing F1-score from 71.8 to 73.4 while deep multimodal methods pushing further to 75.0. This is consistent with previous works indicating that visual contexts are helpful to improve the NER performance.

Secondly, although shallow multimodal methods show a competitive performance in several cases that they get the second best results, they are generally inferior than the deep ones. For example, the shallow multimodal method MAF achieves a best F1-score with 73.42 while the deep multimodal method ICKA achieves 75.02 on the Twitter-15 dataset. This shows that shallow MNER methods suffer from limited uni-modality information and fails in capturing deep interactions across multi-modalities.

Finally, our DESER model achieves new state-of-the-art performance on the two datasets. Specifically, our DESER outperforms the best baseline with 1.01% relative improvement in terms of F1-score on Twitter-15. Furthermore, our DESER gets the best result over all the 10 settings on the two datasets, respectively. The improvement of our DESER over baselines could be attributed to two reasons: i) DESER is a deep stacking of skip-attention layers, which can model the deep and fine-grained interactions between text and image modalities; and ii) The novel skip-attention layer and multimodal gating fusion module are able to debias the unrelated visual contexts to the textual contents by following the text guidance leader and preventing error propagation from noisy image modality during deep multimodal fusing.

Resource-consumption scenario For the training time in resource consumption, our models spend about 102 seconds per epoch using one NVIDIA GTX 3080 Ti GPU. As a reference, it is 83s for the text-only BERT model,

Table 4: Performance comparison with state-of-the-art approaches.

Methods	Twitter-2015							Twitter-2017						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	**LOC**	**ORG**	**MISC**	**Pre**	**Rec**	**F1**	**PER**	**LOC**	**ORG**	**MISC**	**Pre**	**Rec**	**F1**
CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	87.16	80.37
BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
VG	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
ACoA	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
UMT	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
MAF	84.67	81.18	63.35	41.82	71.86	75.10	73.42	91.51	85.80	85.10	68.79	86.16	86.39	86.25
RpBERT	86.36	81.75	62.86	42.40	71.86	76.38	74.07	91.56	84.27	83.54	67.12	84.86	86.31	85.58
HVPNet	86.07	82.15	61.27	42.45	73.11	75.73	74.40	92.13	84.51	84.42	70.78	85.72	87.12	86.42
ITA	86.13	82.70	63.42	41.23	74.42	75.21	74.81	90.78	79.66	80.05	57.78	82.40	83.20	82.80
UMGF	84.26	83.17	62.45	42.42	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51
ICKA	86.54	83.48	64.12	**44.83**	72.15	77.15	75.02	92.45	85.27	86.02	**73.58**	85.07	**88.49**	86.64
DESER	**87.12**	**83.96**	**64.47**	43.53	**75.31**	**77.73**	**75.89**	**93.87**	**86.03**	**86.55**	71.76	**86.89**	87.94	**87.38**

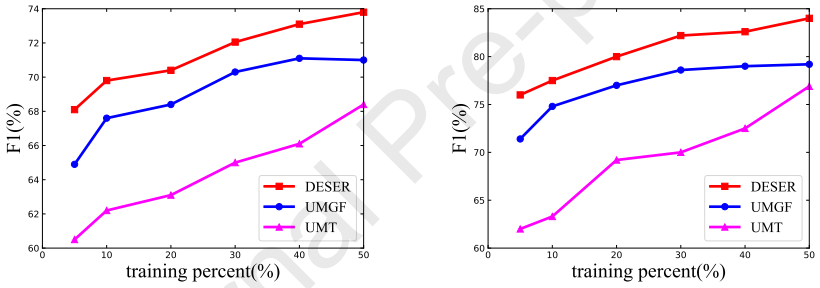


Figure 4: **Performance comparison in low-resource scenario. Left: Twitter-15, Right: Twitter-17**.

125s for the shallow multimodal UMT method, and 217s for the deep multimodal UMGF method. This indicates that our model saves a lot of computing resources while achieving better performance.

Noisy-datasets scenario DESER achieved good results on full scenario. We report the results on the noisy datasets and show that DESER indeed gets improvements in the hard, noisy scenario. To evaluate quantitatively, the two noisy datasets (Twitter-2015-noise and Twitter-2017-noise) are firstly constructed by reserving text-image pairs which have a *Text-Image Relation Classification (TRC)* score less than 0.5. TRC score is used to measure the relevance of text-image pairs [38]. The results on noisy scenario are shown in Table 7 and Table 8. Note, the text-only methods are agnostic to such changes since they do not exploit image information.

Firstly but not surprisingly, all multimodal methods face challenge on

this hard, noisy scenario, since they all drop the performance a lot on both datasets, especially on the Twitter-2015-noise dataset. Notably, five out of the six MNER baselines are generally inferior to the text-only NER methods in this hard scenario.

Secondly, our DESER still outperforms all six MNER methods on both datasets in terms of all evaluation metrics. Specifically, DESER outperforms the best baseline with 2.59% relative improvement in terms of Recall on Twitter-2017-noise dataset. Such improvement we shown that DESER indeed achieves improvements in the noisy scenario.

Low-resource scenario We conduct experiments in low-resource setting by randomly sampling 5% to 50% from the full training set to construct a low-resource training set. The results are shown in Figure 4 where two representative shallow and deep MNER methods are compared. Firstly, we can see that all methods get better performance with the increasing of training set. Secondly, our DESER achieves much advantage under extreme low-resource scenario. In detail, relative improvements of DESER over the shallow UMT method are 22.92% and 8.67% at training percent 5% and 50% respectively, while 7.35% and 6.44% over the deep UMGF method on Twitter-17 dataset. The same trends are observed on Twitter-15. It shows that the neural architecture of DESER is effective in incorporating visual information to complement the text content under data scarcity scenario.

Cross-domain scenario We conduct experiments in cross-domain scenario for transferable generalization analysis. The "Twitter-17 $\rightarrow$ Twitter-15" setting denotes that we train the model on Twitter-17 and then test on Twitter-15 and vice versa. The results are shown in Table 5. Firstly, we can see that all methods degrade performance in cross-domain scenario. Secondly, our DESER significantly outperforms both of the two representative shallow and deep MNER methods by a larger margin. For example, DESER outperforms the deep UMGF method by 3.01% and 2.85% on the two settings, respectively. It shows that DESER architecture has the advantage of transferable generalization, though data distributions are different between the two datasets.

Cross-task scenario To further examine the generalization ability of DESER beyond domain-specific MNER datasets, we additionally conduct cross-task evaluation on the Multimodal Relation Extraction (MRE) dataset. As shown in Table6, our DESER achieves the best overall F1 score among all compared methods, outperforming strong multimodal baselines such as UMGF and HVPNet. This improvement can be attributed to the deep skip-attention

Table 5: Statistics of the two benchmark datasets.

Methods	Twitter-17→ Twitter-15			Twitter-15→Twitter-17		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
UMT	64.67	63.59	64.13	67.80	55.23	60.87
UMGF	67.00	62.81	66.21	69.88	56.92	62.74
DESER	**70.53**	**64.62**	**68.20**	**70.17**	**60.10**	**64.53**

Table 6: Performance comparison on MRE dataset.

Method	MRE		
	Overall Precision	Overall Recall	Overall F1
ACoA	64.67	57.98	61.14
UMT	62.93	63.88	63.46
UMGF	64.38	66.23	65.29
HVPNet	83.64	80.78	81.85
DESER	83.23	83.44	83.76

architecture, which enables effective cross-modal interaction while preserving dominant textual semantics, as well as the multimodal gating mechanism that suppresses irrelevant visual noise.These results demonstrate that DESER generalizes well beyond the MNER task and is applicable to other multimodal extraction tasks, supporting the robustness and transferability of the proposed framework.

4.4. Ablation Study: RQ2

We have shown the effectiveness of DESER comparing with SOTA methods, and in this section, we investigate the performance of skip-attention network and multimodal fusion gating in DESER.

Impact of skip-attention network The Skip-Attention layer(SA) in DESER to model the interactions across modalities and preserve the important information with text guidance by cascading in depth. As shown in Figure 5, DESER w/o SA performs worse than DESER. For example, Precision reduces 3.75% on Twitter-15, and 2.61% on Twitter-17. It verifies that the skip-attention layer is important to predict the entity types and boost the recognition performance with the text guidance.

Impact of multimodal fusion gating The multimodal Gating Fusion module (GF) in DESER adaptively distinguishes dominant modalities from all modalities and controls the contributions of the unrelated modalities by utilizing self-optimizing pseudo-label training. As shown in Figure 5, DESER w/o GF performs worse than DESER. For example, the relative reduction is

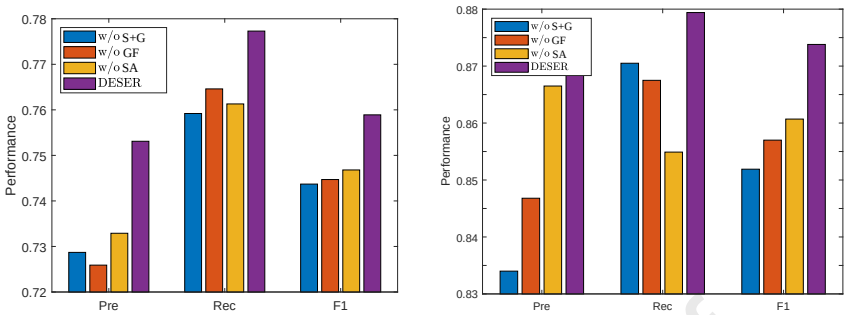


Figure 5: **Impact of multimodal interaction effectiveness (left: Twitter-15, right: Twitter-17), where SA denotes skip-attention, GF denotes multimodal gating fusion, S+G denotes skip-attention + multimodal gating fusion.**

Table 7: Performance on the Twitter-2015-noise dataset.

Methods	Single Type (F1)				Overall		
	PER	**LOC**	**ORG**	**MISC**	**Pre**	**Rec**	**F1**
BERT-CRF	**85.63**	79.95	57.33	33.92	70.11	72.91	71.35
ITA	81.48	77.98	52.92	22.17	68.49	67.50	67.99
UMT	81.24	78.70	50.00	22.69	68.22	67.06	67.64
MAF	82.56	77.50	55.81	31.60	68.66	70.51	69.58
UMGF	84.86	77.61	51.16	24.79	67.55	70.10	68.80
HVPNet	83.86	79.07	53.79	27.32	70.07	69.56	69.82
RpBERT	83.60	79.19	57.32	38.48	70.36	73.27	71.79
DESER	85.23	**79.96**	**58.78**	**39.10**	**71.92**	**73.60**	**72.86**

1.62% and 1.52% in terms of F1 metric on Twitter-15 and Twitter-17 respectively. It shows that it is necessary to debias the visual contexts when the associated image is unrelated to the text and to prevent the error propagation from the noisy image modality during deep multimodal fusing.

Obviously, DESER w/o S+G is the worst where Precision reduces 4.18% and F1 reduces 2.57% on Twitter-17. It further shows that both of skip-attention layer and multimodal Gating Fusion module in DESER contribute to the performance improvements.

4.5. Case Study: RQ3

We have quantitatively show the effectiveness of DESER in both full (relevant+noisy) and noisy scenarios. In this section, we illustrate the results of DESER and baselines on representative cases. The first group is relevant

Table 8: Performance on the Twitter-2017-noise dataset.

Methods	Single Type (F1)				Overall		
	PER	**LOC**	**ORG**	**MISC**	**Pre**	**Rec**	**F1**
BERT-CRF	90.96	**85.96**	81.90	65.84	84.67	85.05	84.86
ITA	89.99	75.84	80.21	62.54	81.94	81.57	81.76
UMT	90.02	81.19	80.19	64.44	83.13	83.95	83.54
MAF	90.60	78.24	82.55	66.04	83.20	84.14	83.67
UMGF	90.25	83.85	80.19	63.80	81.67	84.75	83.16
HVPNet	90.21	83.78	81.12	66.23	83.00	84.90	83.94
RpBERT	91.98	82.66	83.42	66.23	85.36	85.42	85.39
DESER	**92.32**	84.43	**84.83**	**68.51**	**85.49**	**87.64**	**86.75**

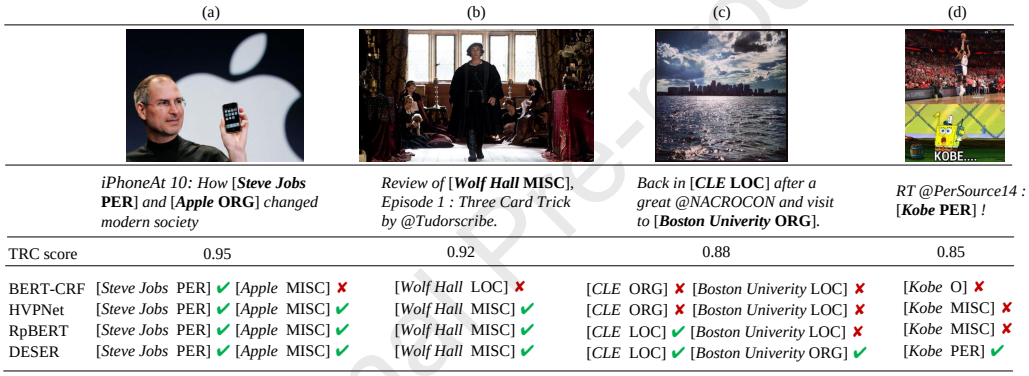


Figure 6: Case study I: On relevant text-image pairs (TRC score is greater than 0.5).

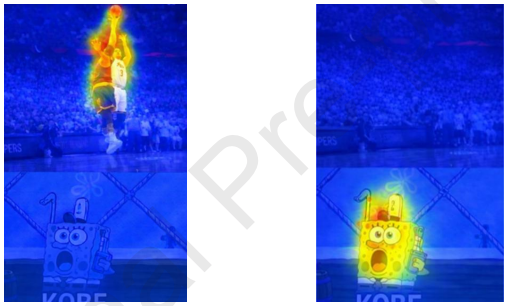
text-image cases while the second is noisy cases.

Case study I: On relevant text-image pairs As shown in Figure 6, the MNER methods are generally better than the text-only NER methods, especially when the text-image pairs are highly relevant (TRC score greater than 0.9 in Figure 6(a) and (b)). Furthermore, the two MNER baselines have some difficulty in Figure 6 (c) and (d). They are wrong in recognizing "Kobe" due to the alignment between visual object and text entity is hard to disclose. This shows that DESER can attend the most relevant visual region in image by fuse text and image representations in an effective way.

Case study II: On noisy text-image pairs As shown in Figure 7, the text-only NER method shows a strong performance in noisy scenario since the MNER baselines may introduce the noise from image into the text modality. With the noise going worse (smaller TRC score), baselines are getting deteriorated. Our DESER is much impressive in such hard, noisy cases.

	(a)	(b)	(c)	(d)
				
	[*Brad Marchand* **PER**]	*Ask* [***Siri* MISC**] *what 0 divided by 0 is and watch her put you in your place.*	*Nice image of* [***Kevin Love* PER**] [1] *and* [***Kyle Korver* PER**] [2] *during 1st half # NBAFinals # Cavsin9 #* [***Cleveland* ORG**] [3]	*Me trying to explain* [***Harry Potter* MISC**] *to a muggle*
TRC score	0.35	0.28	0.17	0.17
BERT-CRF	[*Brad Marchand* PER] ✔	[*Siri* MISC] ✔	[1 PER]✔ [2 PER]✔ [3 LOC]✘	[*Harry Potter* PER] ✘
HVPNet	[*Brad Marchand* MISC] ✘	[*Siri* PER] ✘	[1 MISC]✘ [2 MISC]✘ [3 LOC]✘	[*Harry Potter* PER] ✘
RpBERT	[*Brad Marchand* MISC] ✘	[*Siri* MISC] ✔	[1 PER]✔ [2 PER]✔ [3 ORG]✔	[*Harry Potter* PER] ✘
DESER	[*Brad Marchand* PER] ✔	[*Siri* MISC] ✔	[1 PER]✔ [2 PER]✔ [3 ORG]✔	[*Harry Potter* MISC] ✔

Figure 7: Case study II: On noisy text-image pairs (TRC score is less than 0.5).



(a) DESER (b) HVPNet

Figure 8: The attention visualization on "Kobe".

Note, baseline RpBERT is better than the baseline HVPNet in noisy scenario since it adopts an auxiliary task to compute the relation score between a pair of text and image. In contrast, our DESER neither trains on external sources nor learns multi-tasks. DESER attacks the noisy issue by an effective DESER network, which shows the multimodal gating fusion network reduces the influence of noisy modalities information and the pseudo-label training controls the contributions of different modalities to decision by self-optimizing adaptively.

Case study III: Attention visualization of skip-attention. To further illustrate how the proposed skip-attention mechanism operates across modalities, Figure8 visualizes how different models attend to image regions when processing the same multimodal input associated with the entity "Kobe". As shown in Figure8(a), DESER with skip-attention selectively focuses on the semantically relevant visual regions (i.e., the athlete in the upper im-

age), while effectively suppressing irrelevant or misleading content in the lower part of the image. In contrast, Figure 8(b) shows that the baseline model (HVPNet) assigns high attention to visually salient but semantically irrelevant regions, indicating weaker cross-modal alignment.

5. DISCUSSION

In our approach to Multimodal Named Entity Recognition (MNER), the representation of visual information plays a pivotal role in bridging the semantic gap between text and images. We adopt ResNet as the image encoder due to its strong capability to extract deep semantic features from images. Rather than using a single global image representation, we follow a spatial decomposition strategy by dividing each input image into $7 \times 7 = 49$ local visual blocks. Each block is independently encoded into a 2048-dimensional feature vector by ResNet, enabling the model to capture localized visual cues that are often crucial for fine-grained entity recognition. To align these visual representations with the textual modality, we project the 2048-dimensional image features into a common semantic space of dimensionality d using a learned linear transformation. This mapping ensures compatibility with BERT-derived textual embeddings and allows subsequent attention mechanisms to perform effective multimodal fusion. Importantly, such alignment allows the model to not only co-attend across modalities but also dynamically prioritize visual regions that are semantically relevant to named entities in the text. The adoption of region-level features rather than holistic image embeddings is particularly beneficial in the social media domain, where images often contain multiple elements—some of which may be irrelevant to the accompanying text. This fine-grained design enables the DESER model to leverage skip-attention layers that selectively focus on important visual blocks under the guidance of textual queries, preserving critical textual information while incorporating complementary visual context. However, this strategy introduces certain limitations. The fixed-grid partitioning may not always align with semantically coherent regions, potentially introducing visual noise. To mitigate this, DESER integrates a multimodal gating fusion module, which adaptively adjusts the influence of the image modality based on learned pseudo-labels. This mechanism proves especially effective in noisy or ambiguous contexts, helping to suppress misleading visual signals and amplify the more reliable text modality. Overall, the integration of ResNet-based spatial image encoding, semantic projection, and adaptive

fusion within the DESER framework illustrates a robust pathway for achieving enhanced multimodal interaction and mitigating modality-specific biases in MNER. This careful balance of image granularity and modality control contributes significantly to the model's state-of-the-art performance across various experimental settings.

6. CONCLUSION

In this paper, we proposed a Deep Modular Skip-Attention Network for Multimodal Named Entity Recognition (DESER), to address two major challenges in existing multimodal systems: ineffective deep interaction modeling and vulnerability to noisy visual content. By integrating a novel Skip-Attention mechanism and a multimodal gating fusion module with pseudo-label training, DESER enables fine-grained, text-guided cross-modal interactions while dynamically mitigating the influence of irrelevant or misleading visual information.

Our experiments across multiple scenarios-including full, noisy, low-resource, and cross-domain settings-demonstrate that DESER consistently outperforms state-of-the-art models on two widely-used benchmark datasets. The results confirm that DESER not only improves multimodal interaction effectiveness but also enhances model robustness and generalizability.

Our main contributions can be summarized as follows: (i) Skip-Attention facilitates deeper multimodal stacking with preserved textual semantics and reduced computation; (ii) the gating fusion with pseudo-labeling provides adaptive modality weighting, effectively debiasing visual noise; and (iii) the overall architecture supports scalability and transferability.

In future work, we plan to explore task-agnostic pretraining of the skip-attention layers and extend DESER to other multimodal tasks such as relation extraction and event detection. Our findings contribute a strong foundation for robust and efficient multimodal learning in noisy, real-world environments.

CRediT authorship contribution statement

Chengen Lai: Conceptualization, Methodology, Writing - original draft, Software. **Shengli Song**: Writing - review & editing, Validation, Supervision. **Shiqi Meng**: Data curation, Formal analysis. **Guangneng Hu**: Visualization, Investigation. Kun Fu: Supervision.

References

[1] L. Sun, J. Wang, Y. Su, F. Weng, Y. Sun, Z. Zheng, Y. Chen, Riva: a pre-trained tweet multimodal model based on text-image relation for multimodal ner, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1852–1862.

[2] S. Moon, L. Neves, V. Carvalho, Multimodal named entity recognition for short social media posts, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 852–860.

[3] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, Visual attention model for name tagging in multimodal social media, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1990–1999.

[4] Q. Zhang, J. Fu, X. Liu, X. Huang, Adaptive co-attention network for named entity recognition in tweets, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, 2019, pp. 4171–4186.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[7] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-f. Leung, Q. Li, Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1038–1046.

[8] B. Xu, S. Huang, C. Sha, H. Wang, Maf: a general matching and alignment framework for multimodal named entity recognition, in: Proceedings of the fifteenth ACM international conference on web search and data mining, 2022, pp. 1215–1223.

[9] J. Yu, J. Jiang, L. Yang, R. Xia, Improving multimodal named entity recognition via entity span detection with unified multimodal transformer, Association for Computational Linguistics, 2020.

[10] X. Chen, N. Zhang, L. Li, Y. Yao, S. Deng, C. Tan, F. Huang, L. Si, H. Chen, Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 1607–1618.

[11] L. Sun, J. Wang, K. Zhang, Y. Su, F. Weng, Rpbert: a text-image relation propagation-based bert model for multimodal ner, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 13860–13868.

[12] X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, K. Tu, Ita: Image-text alignments for multi-modal named entity recognition, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 3176–3189.

[13] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, G. Zhou, Multi-modal graph fusion for named entity recognition with targeted visual guidance, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 14347–14355.

[14] G. Luo, X. Huang, C.-Y. Lin, Z. Nie, Joint entity recognition and disambiguation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 879–888.

[15] F. Li, Z. Wang, S. C. Hui, L. Liao, D. Song, J. Xu, Effective named entity recognition with boundary-aware bidirectional neural networks, in: Proceedings of the Web Conference 2021, 2021, pp. 1695–1703.

[16] Y. Tian, X. Sun, H. Yu, Y. Li, K. Fu, Hierarchical self-adaptation network for multimodal named entity recognition in social media, Neurocomputing 439 (2021) 12–21.

[17] F. Wu, J. Liu, C. Wu, Y. Huang, X. Xie, Neural chinese named entity recognition via cnn-lstm-crf and joint training with word segmentation, in: The World Wide Web Conference, 2019, pp. 3342–3348.

[18] Y. Zhou, L. Huang, T. Guo, S. Hu, J. Han, An attention-based model for joint extraction of entities and relations with implicit entity features, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 729–737.

[19] A. Iovine, A. Fang, B. Fetahu, O. Rokhlenko, S. Malmasi, Cyclener: an unsupervised training approach for named entity recognition, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2916–2924.

[20] X. Zhang, B. Yu, Y. Wang, T. Liu, T. Su, H. Xu, Exploring modular task decomposition in cross-domain named entity recognition, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 301–311.

[21] C. Mai, M. Qiu, K. Luo, Z. Peng, J. Liu, C. Yuan, Y. Huang, Pretraining multi-modal representations for chinese ner task with cross-modality attention, in: Proceedings of the fifteenth ACM international conference on web search and data mining, 2022, pp. 726–734.

[22] X. Wang, J. Tian, M. Gui, Z. Li, J. Ye, M. Yan, Y. Xiao, Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition, in: International Conference on Database Systems for Advanced Applications, Springer, 2022, pp. 297–305.

[23] C. Zheng, Z. Wu, T. Wang, Y. Cai, Q. Li, Object-aware multimodal named entity recognition in social media posts with adversarial learning, IEEE Transactions on Multimedia 23 (2020) 2520–2532.

[24] J. Li, H. Li, D. Sun, J. Wang, W. Zhang, Z. Wang, G. Pan, Llms as bridges: Reformulating grounded multimodal named entity recognition, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 1302–1318.

[25] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, p. 6558.

[26] D. Chen, Z. Li, B. Gu, Z. Chen, Multimodal named entity recognition with image attributes and image knowledge, in: International Conference on Database Systems for Advanced Applications, 2021, pp. 186–201.

[27] X. Wang, J. Ye, Z. Li, J. Tian, Y. Jiang, M. Yan, J. Zhang, Y. Xiao, Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 1–6.

[28] Q. Zeng, M. Yuan, J. Wan, K. Wang, N. Shi, Q. Che, B. Liu, Icka: an instruction construction and knowledge alignment framework for multimodal named entity recognition, Expert Systems with Applications 255 (2024) 124867.

[29] Z. Li, J. Yu, J. Yang, W. Wang, L. Yang, R. Xia, Generative multimodal data augmentation for low-resource multimodal named entity recognition, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7336–7345.

[30] J. Lu, D. Zhang, J. Zhang, P. Zhang, Flat multi-modal interaction transformer for named entity recognition, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 2055–2064.

[31] M. Jia, L. Shen, X. Shen, L. Liao, M. Chen, X. He, Z. Chen, J. Li, Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 8032–8040.

[32] F. Zhao, C. Li, Z. Wu, S. Xing, X. Dai, Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3983–3992.

[33] E. T. K. Sang, J. Veenstra, Representing text chunks, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999, pp. 173–179.

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[35] Q. Meng, S. Zhao, Z. Huang, F. Zhou, Magface: A universal representation for face recognition and quality assessment, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14225–14234.

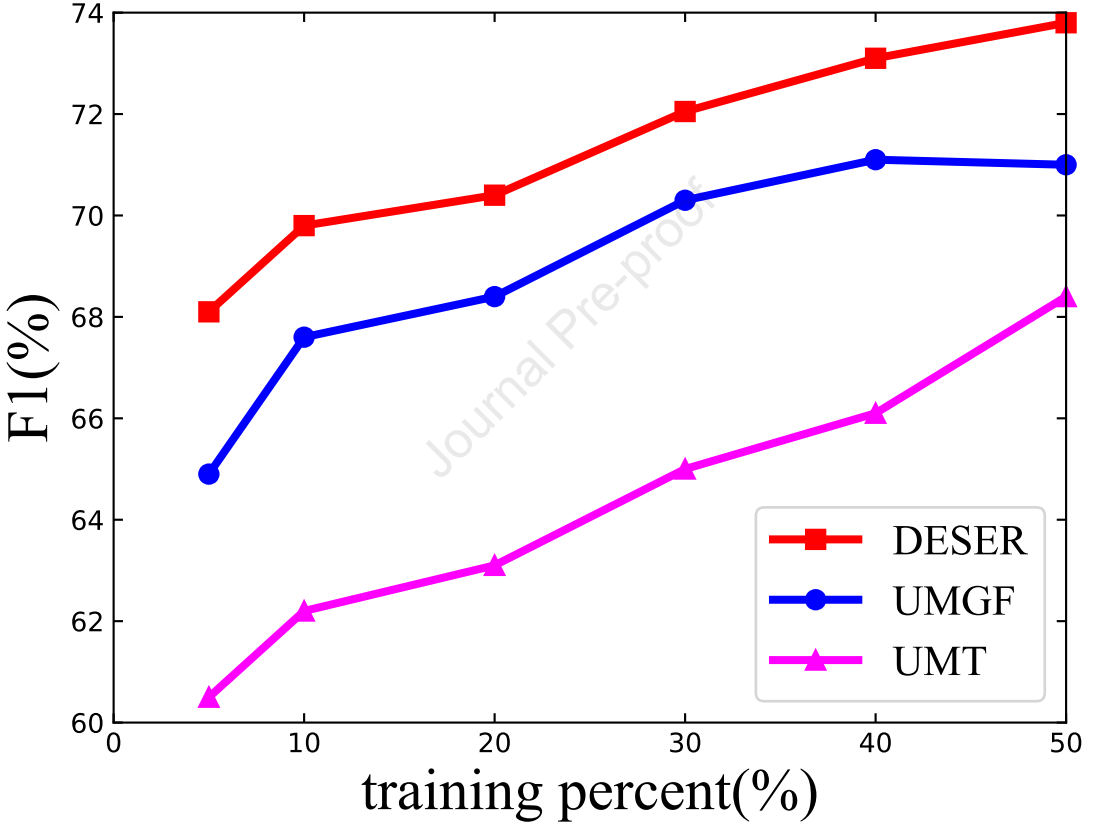
[36] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1064–1074.

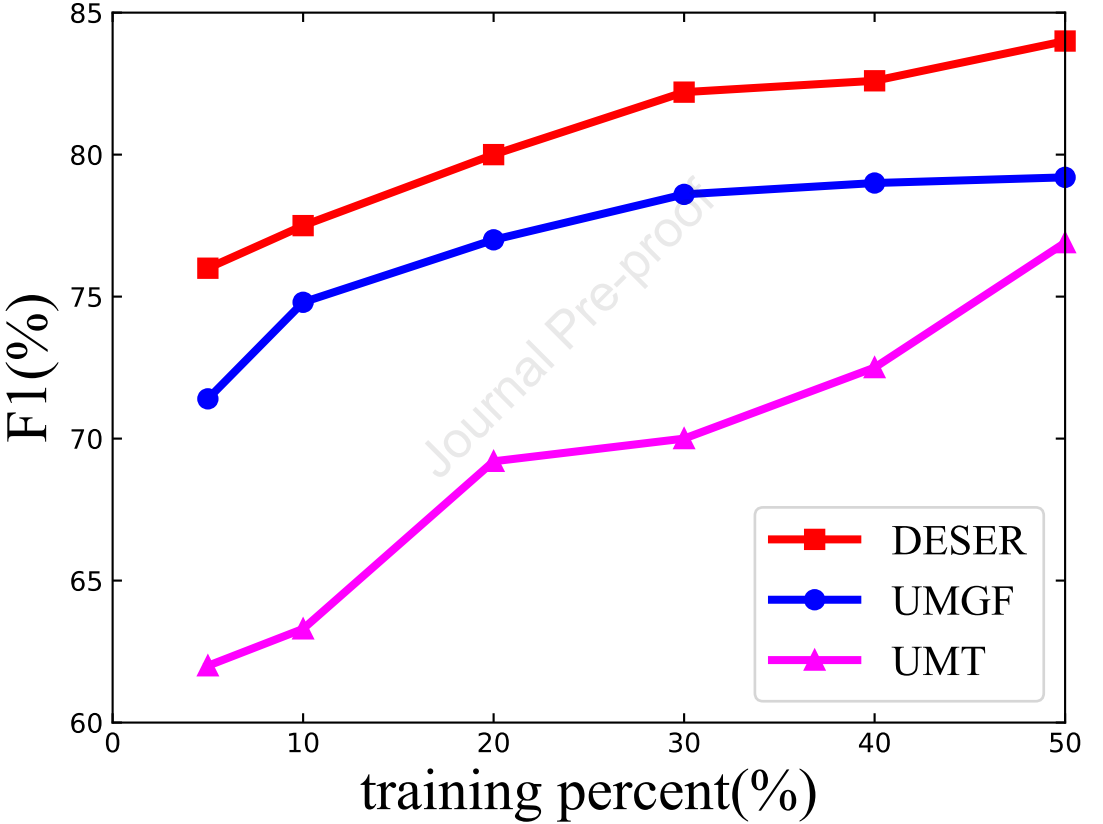
[37] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 260–270.

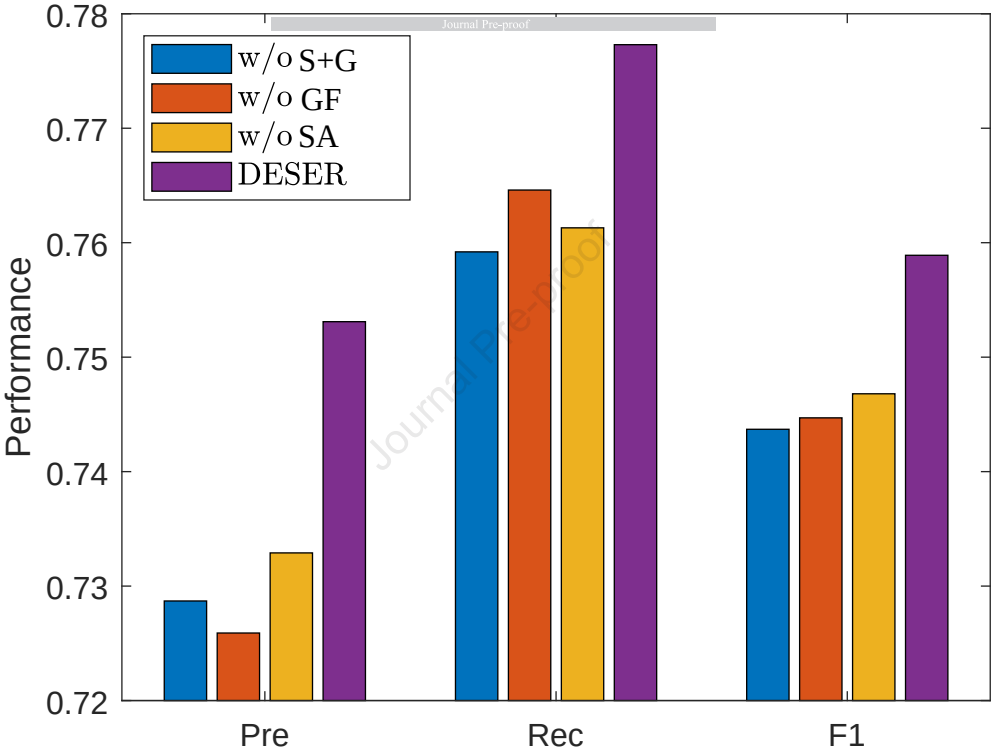
[38] A. Vempala, D. Preoţiuc-Pietro, Categorizing and inferring the relationship between the text and image of twitter posts, in: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, 2019, pp. 2830–2840.

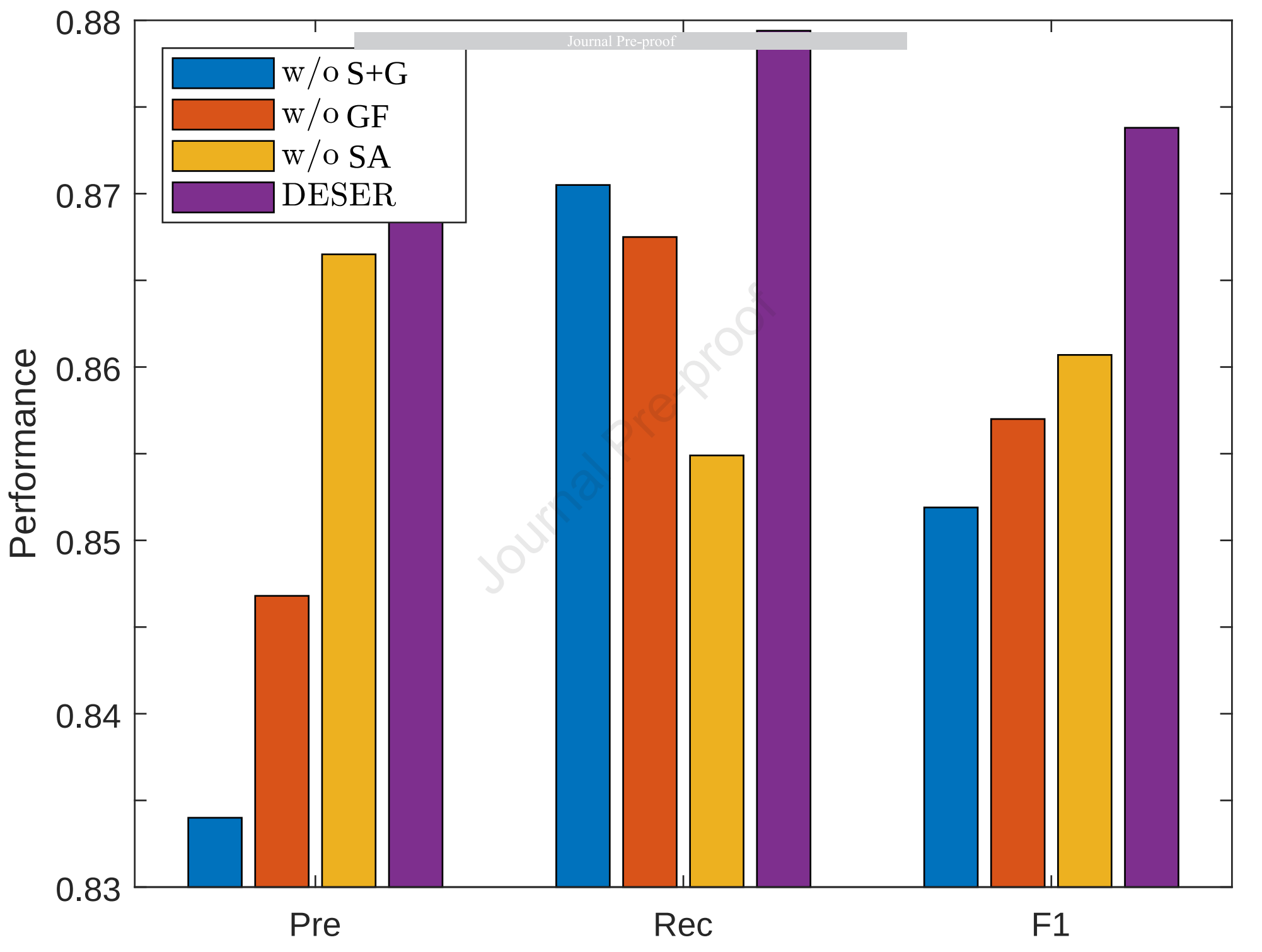
Acknowledgments

This work is supported by National Natural Science Foundation of China (62306220).









Chengen Lai obtained his Bachelor's degree in computer science and technology from Xidian University, Xi'an, China, in 2020 and the Ph.D. degree from Xidian University, Xi'an, China in 2025. He is currently a senior algorithm engineer at Alibaba's Tongyi Lab.

Shengli Song received the B.E. degree from Xidian University, Xi'an, China, in 2003, the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2006 and the Ph.D. degree from Xidian University, Xi'an, China in 2010. He is currently a Professor in the School of Computer Science and Technology at Xidian University. His research interests are in cognitive computing, natural language processing, and large language model.

Sitong Yan obtained her Bachelor's degree in computer science and technology from Xidian University, Xi'an, China, in 2020. Currently, she is a Ph.D. candidate at the School of Computer Science and Technology, Xidian University. Her current research interests include mixed-initiative interaction, dialogue system and natural language processing.

Guangneng Hu received the B.E. degree from Nanjing University in 2013, the M.S. degree from Nanjing University in 2016 and the Ph.D. degree from Hong Kong University of Science and Technology in 2020, China. He is currently an Associate Professor in the School of Computer Science and Technology at Xidian University. His research interests are in machine learning and natural language processing, etc. He has long served as the program committee member of top international journals and conferences, and the program chairman of the AAAI 2021 organizing committee.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: