Boosting Textural NER with Synthetic Image and Instructive Alignment

Jiahao Wang[1] , **Wenjun Ke**[1,2*] , **Peng Wang**[1,2*] , **Hang Zhang**[3] ,
Dong Nie[3†] , **Jiajun Liu**[1] , **Guozheng Li**[1] , **Ziyu Shang**[1]
[1]School of Computer Science and Engineering, Southeast University
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
[3]Beijing Institute of Computer Technology and Application, Beijing
[4]Alibaba Inc. US
{wang_jh, kewenjun, pwang, jiajliu, gzli,ziyus1999}@seu.edu.cn,
zhanghangjkluio@163.com, dongnie@cs.unc.edu

Abstract

Named entity recognition (NER) is a pivotal task reliant on textual data, often impeding the disambiguation of entities due to the absence of context. To tackle this challenge, conventional methods often incorporate images crawled from the internet as auxiliary information. However, the images often lack sufficient entities or would introduce noise. Even with high-quality images, it is still challenging to achieve fine-grained alignment with texts. We introduce a novel method named InstructNER to address these issues. Leveraging the rich real-world knowledge and image synthesis capabilities of a large pre-trained stable diffusion model, InstructNER transforms the text-only NER into a multimodal NER (MNER) task. A selection process automatically identifies the best synthetic image by comparing fine-grained similarities with internet-crawled images through a visual bag-of-words strategy. Note, during the image synthesis, a cross-attention matrix between synthetic images and raw text emerges, which inspires a soft attention guidance alignment (AGA) mechanism. AGA optimizes the MNER task and concurrently facilitates instructive alignment in MNER. Experiments on prominent MNER datasets show that our method surpasses all text-only baselines, improving F1-score by 1.4% to 2.3%. Remarkably, even when compared to fully multimodal baselines, our approach maintains competitive. Furthermore, we open-source a comprehensive synthetic image dataset and the code to supplement existing raw dataset. The code and datasets are available in https://github.com/Heyest/InstructNER.

1 Introduction

Named entity recognition (NER) is a fundamental information extraction task that identifies named entities in sentences and classifies them into predefined categories(Li et al., 2020). Text-based NER methods in practice mainly depend on textual information, which poses challenges in recognizing short or ambiguous sentences (Zhang et al., 2018). In comparison with text-based NER, multimodal NER (MNER) leverages images as supplementary information to boost the task's robustness (Xu et al., 2022; Jia et al., 2022; Wang et al., 2023a). In Figure 1 (a), the word *ICRC* in the sentence *I love ICRC* is ambiguous, hindering the classification based solely on the text. An additional image clarifies that *ICRC* refers to the International Committee of the Red Cross, labeled as *ORG*.

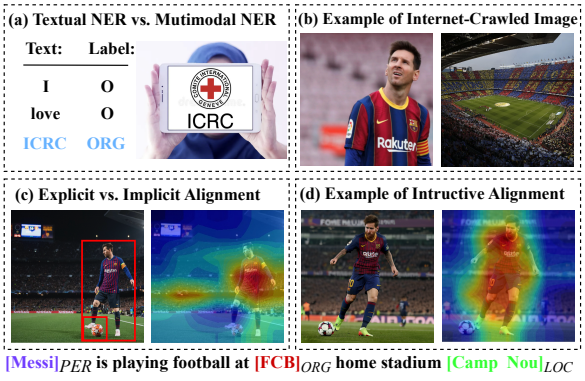

Figure 1: Examples of NER and MNER, where *PER*, *ORG*, *LOC* and *MISC* denote the entity labels of person, organization, location and others, respectively.

Despite the promising results of MNER methods (Chen et al., 2021), they still struggle with acquiring large-scale and high-quality text-image paired corpus. Common MNER datasets, Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018), consist of 8,257 and 4,819 image-text pairs, respectively. This represents a mere 26% and 16% of the volume found in the frequently employed textual NER dataset CoNLL03 (Tjong Kim Sang and De Meulder, 2003). One potential solution involves crawling supplementary images by internet search engines. Whereas, unlike raw texts, the internet-crawled images often fail to cover enough entities,

*Corresponding author.
†Now working in Meta.

Findings of the Association for Computational Linguistics ACL 2024, pages 1277–1287

which might further bring unexpected noise. As exemplified in Figure 1 (b), the two images are retrieved from Google based on *Messi is playing football at FCB home stadium Camp Nou*, where the left image only contains *Messi* while the right one only includes *Camp Nou*.

Another critical challenge to the success of the MNER task lies in the accurate alignment between text tokens and image regions. Existing techniques for the text-image alignment can be summarized into two groups: explicit alignment and implicit alignment. Explicit alignment extracts visual object regions, then maps them to corresponding textual tokens (Wu et al., 2020; Zheng et al., 2020; Jia et al., 2022), but errors can propagate from initial inaccurate region extraction (Yang et al., 2019a). Consider the left image in Figure 1 (c), only the object regions of *Messi* and *football* are extracted, missing the *FCB* clothes logo and *Camp Nou* stadium in the background. In contrast, implicit alignment alleviates this issue by employing an attention mechanism to learn the alignment weights adaptively (Zhou et al., 2022; Xu et al., 2022; Wang et al., 2023a). Despite the token-to-image alignment achieved by these approaches, the attention may not be sufficiently concentrated. Regarding the right image of Figure 1 (c), the attention heatmap illustrates *Messi* is located, yet it noticeably allocates irrelevant attention to surrounding areas.

In response to the aforementioned challenges, we propose InstructNER, a novel approach that harnesses the rich real-world knowledge and the image synthesis capabilities of stable diffusion models (Rombach et al., 2022) to provide supplemental information, thereby transforming the textual NER task into MNER task. Specifically, we first feed the raw text into the pre-trained stable diffusion (SD) model to generate large-scale synthetic images. Then, to alleviate the variability of image quality and select the best one, we employ an off-the-shelf visual bag-of-words (BoW) method (Gidaris et al., 2020) to measure the fine-grained similarity between the internet-crawled images and the synthetic images, ultimately selecting the most similar one. Particularly, the similarity metric of images considers both the coverage and accuracy of entities. As depicted in Figure 1 (d) (left), the selected image accurately encompasses all entities, including *Messi*, the *FCB* logo, and the *Camp Nou* stadium. Furthermore, the image synthesis process also produces a cross-attention matrix between the synthetic images and the raw text as a byproduct, inspiring our soft attention guidance alignment (AGA) mechanism for MNER model training. In addition to optimizing the original MNER task, our objective also aims to minimize the Kullback-Leibler (KL) divergence between the MNER model's cross-attention matrix and the aforementioned byproduct matrix. Figure 1 (d) shows the attention of the AGA mechanism paid to token *Messi*, displaying a higher concentration.

We conduct experiments on three representative MNER datasets, $i.e.$, Twitter-2015, Twitter-2017, and WikiDiverse, while excluding images with only textual corpus. Experimental results demonstrate the superiority of our method over all text-only baselines, with absolute F1-score improvements of 1.4%, 2.3% and 2.1%. Moreover, our method still achieves competitive results even when compared to fully multimodal baselines. To sum up, the contributions of this paper are three-fold:

- We are the first to leverage the artificial intelligence generated content (AIGC) ability of stable diffusion model to switch textual NER into MNER with synthetic images.
- We propose a comprehensive framework InstructNER with a novel text-to-image mechanism AGA, which instructs the cross-attention being learned in a soft manner.
- Experimental results compared with both text-only and fully MNER baselines verify the effectiveness of our method. Moreover, we release the large-scale and high-quality synthetic images to supplement the raw datasets.

2 Method

NER aims to categorize named entities in a sentence $S = (s_1, s_2, \ldots, s_n)$ consisting of n tokens, and often adopts the BIO tagging schema (Sang and Veenstra, 1999). The output $Y = (y_1, y_2, \ldots, y_n)$ consists of n labels, where $y_i \in \mathcal{L}$ and $\mathcal{L} = \{\mathcal{B}, \mathcal{I}, \mathcal{O}\}$ represents the predefined label set. The MNER task receives an image as additional input, then identifies named entities similar to NER.

Figure 2 provides an overview of the comprehensive architecture of InstructNER. In Stage #1, InstructNER leverages the pre-trained stable diffusion model to generate images, and selects the optimal synthetic image covering sufficient and accurate entities. In Stage #2, InstructNER utilizes the soft attention guidance alignment (AGA) mechanism to fuse the raw text and synthetic image. In

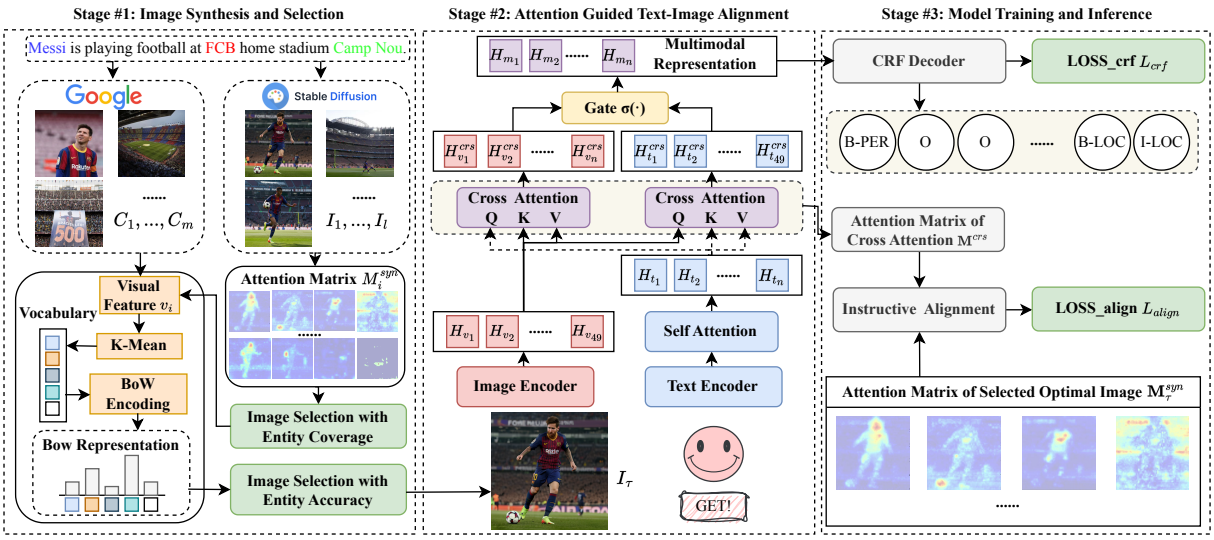

Figure 2: Illustration of InstructNER. The arrows indicate the data flow from the previous stage to the next one.

Stage #3, InstructNER adopts a multi-task training object to cocurrently optimize the loss of entity classification and instructive alignment.

2.1 Image Synthesis and Selection

For the sentence $S = (s_1, s_2, \ldots, s_n)$, we first generate l synthetic images $I = \{I_1, I_2, \ldots, I_l\}$ using the pre-trained SD model (Rombach et al., 2022), and then select the best one $I_\tau \in I$ considering the entity coverage and accuracy.

2.1.1 Image Selection with Entity Coverage

For each synthetic image $I_i \in I$, we derive its cross attention matrix $\mathbf{M}_i^{syn}$ from pre-trainted SD model, where $\mathbf{M}_i^{syn} = (\mathbf{m}_{i,1}, \mathbf{m}_{i,2}, \ldots, \mathbf{m}_{i,n})$, and $\mathbf{m}_{i,j} \in \mathbb{R}^{d_p}$ represents the attention score vector between token s_j and image I_i (decomposed into d_p pixels). Notably, the entity tokens ($e.g.$, *Messi* and *Camp Nou*) could receive higher attention scores, while function tokens ($e.g.$, *is* and *at*) obtain lower attentions. For each token s_j, its importance w_j is measured by summing up the average attention score $\mu_{i,j}$, which corresponds to all the image pixels:

$$\mu_{i,j} = \frac{1}{d_p} \sum_{k=1}^{d_p} \mathbf{m}_{i,j}[k], w_j = \sum_{i=1}^{l} \mu_{i,j} \quad (1)$$

We use a predefined threshold θ to select the entity tokens. In particular, $w_j \geq \theta$ indicates that token s_j is the desired entity token. We can then filter out the images which exhibit low attention scores with the entity tokens. To this end, we take this strategy to rank the quality of the synthetic image set I, and keep α best images among them, denoted as I', where $I' \subset I$ and α is a hyper-parameter.

2.1.2 Image Selection with Entity Accuracy

The synthetic image set I' excel in entity coverage, but may include some unrealistic and inaccurate contents. Conversely, internet-crawled images often exhibit high factual but lack sufficient entity coverage. Combing the advantage of these two types of images, we employ internet-crawled images to filter synthetic images and select the optimal image I_τ from I'.

Specifically, we crawl m images $C = \{C_1, C_2, \ldots, C_m\}$ from a search engine. For a specific image $C_i \in I' \cup C$, we utilize the pre-trained RotNet encoder (Gidaris et al., 2018) to obtain visual features $\mathbf{v}_i = (\mathbf{v}_i^1, \mathbf{v}_i^2, \ldots, \mathbf{v}_i^r)$, which consist of r regions, where $\mathbf{v}_i^r \in \mathbb{R}^c$ represents a c-dimensional vector. To measure the fine-grained similarity between C and I', an off-the-shelf visual bag-of-words (BoW) model (Gidaris et al., 2020) is employed. K-Means algorithm is first adopted in clustering all the region features $\{\mathbf{v}_i\}|_{i=1}^{|I' \cup C|}$ within the image set $I' \cup C$ into k-cluster, in order to calculate the image similarity with vocabulary consisting of k vectors. Formally, the vocabulary $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k)$ can be represented as:

$$(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k) = \text{KMeans}\left(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{|I' \cup C|}\right) \quad (2)$$

where $\mathbf{f}_k \in \mathbb{R}^c$ denote a specific vector. Then, we quantize each region of the image C_i with the most similar (the minimum squared Euclidean distance) feature in the vocabulary, and compute the visual bag-of-words representation $bow(C_i)$.

$$bow(C_i) = \left\{bow(C_i^j)\right\}\Big|_{j=1}^{k} \quad (3)$$

$$bow(C_i^j) = \mathbb{1}\left(\mathrm{ED}(\mathbf{f}_j, \mathbf{v}_i) < \eta\right) \quad (4)$$

where $\mathbb{1}[\cdot]$ is the indicator operator, $\mathrm{ED}(\cdot)$ denotes the minimum squared Euclidean distance, η is a threshold manually defined. Following the steps above, we compute the BoW representations for all images in $I' \cup C$, and select the most similar image in I' comparing with C as I_τ. Following the steps above, we compute the BoW representations for all images in $I' \cup C$, and select the most similar image in I' comparing with C as I_τ.

$$I_\tau = \arg\min_{I_i \in I'} \sum_{j=1}^{m} \mathrm{ED}\left(bow(I_i), bow(C_j)\right) \quad (5)$$

where $\mathrm{ED}(\cdot)$ denotes the minimum squared Euclidean distance.

2.2 Attention Guided Text-Image Alignment

After obtaining the optimal synthetic image I_τ, we reframe the text-only NER into MNER, *i.e.*, first performing text and image representation and second conducting instructive multimodal alignment.

2.2.1 Text and Image Encoding

For sentence S, BERT (Devlin et al., 2019) is employed to encode contextualized textual representations $\hat{\mathbf{H}}_t \in \mathbb{R}^{n \times d}$, where n and d denotes the sentence length and hidden dimension. Then, to further guide the interaction within textual modality, a self-attention transformer (Vaswani et al., 2017) is utilized. We first map the initial text representation $\hat{\mathbf{H}}_t$ as query $\mathbf{Q}_t$, key $\mathbf{K}_t$ and value $\mathbf{V}_t$ by different linear projections, where $\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t \in \mathbb{R}^{n \times d}$, then use attention mechanism to obtain the interacted text representations $\mathbf{H}_t \in \mathbb{R}^{n \times d}$.

$$\mathbf{M}_t = \mathrm{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d_k}}\right), \mathbf{H}_t = \mathrm{LN}(\hat{\mathbf{H}}_t + \mathbf{M}_t \mathbf{V}_t) \quad (6)$$

where $\mathbf{M}_t \in \mathbb{R}^{n \times n}$, LN denotes the normalization function of transformer layer . Moreover, for the image I_τ, we resize it to 224 $\times$ 224 pixels and derive the visual output $\mathrm{ResNet}(I_\tau) = \{\mathbf{h}_j | \mathbf{h}_j \in \mathbb{R}^{2048}, j = 1, 2, \ldots, 49\}$ from the last convolutional layer of pre-trained ResNet (He et al., 2016). The obtained output is divided into 49 regions, with each region as a 2048-dimensional vector $\mathbf{h}_j$. To align these visual representations with the textual counterparts, a linear mapping operation with parameter $\mathbf{W}_v \in \mathbb{R}^{d \times 2048}$ is applied, resulting in the final visual representation $\mathbf{H}_v \in \mathbb{R}^{49 \times d}$.

$$\mathbf{H}_v = \mathbf{W}_v \cdot \mathrm{ResNet}(I_\tau) \quad (7)$$

2.2.2 Instructive Text-Image Alignment

After obtaining the unimodal text representation $\mathbf{H}_t$ and image representation $\mathbf{H}_v$ respectively, we further utilize two cross-attention transformers (Yu et al., 2020) to facilitate interaction between modalities. As illustrated in Stage #2, we first takes $\mathbf{H}_v$ as query and $\mathbf{H}_t$ as key and value, learning the image-aware token representation $\mathbf{H}_t^{crs} \in \mathbb{R}^{49 \times d}$.

$$\mathbf{H}_t^{crs} = \text{Cross-ATT}(\mathbf{H}_v, \mathbf{H}_t, \mathbf{H}_t) \quad (8)$$

where Cross-ATT$(\cdot)$ denotes the cross-attention layers. Then, we take $\mathbf{H}_t$ as query and $\mathbf{H}_v$ as key and value, learning the token-aware visual representation:

$$\mathbf{H}_v^{crs} = \text{Cross-ATT}(\mathbf{H}_t, \mathbf{H}_v, \mathbf{H}_v) \quad (9)$$

In order to guide the attention in the cross-attention transformers to be more focused, our AGA mechanism utilize the byproduct of attention matrix $\mathbf{M}_\tau^{syn}$ between each token and image I_τ generated by SD to guide the soft attention alignment. In the above two cross-attention transformers, two attention matrixes can be obtained similar to $\mathbf{M}_t$ in Equation 6. We average the two attention matrix to obtain $\mathbf{M}^{crs} = \{\mathbf{m}_1^{crs}, \mathbf{m}_2^{crs}, \ldots, \mathbf{m}_n^{crs}\}$, where $\mathbf{m}_i^{crs} \in \mathbb{R}^{49}$ represents the attention scores between word s_i and the 49 regions of image I_τ. However, $\mathbf{M}_\tau^{syn} = \{\mathbf{m}_1^{syn}, \mathbf{m}_2^{syn}, \ldots, \mathbf{m}_n^{syn}\}$, where $\mathbf{m}_i^{syn} \in \mathbb{R}^{d_p}$ represents the attention between word s_i and all d_p image pixels. To facilitate comparison, we use an average pooling operation to convert $\mathbf{m}_i^{syn}$ to the same dimension as $\mathbf{m}_i^{crs}$. Finally, we calculate the KL divergence between the two attention matrices and use it as an auxiliary loss.

$$L_{align} = \sum_{i=1}^{n} \mathrm{D_{KL}}\left(q\left(\mathbf{m}_i^{crs}\right) || p(\mathbf{m}_i^{syn})\right) \quad (10)$$

where $q(\cdot)$ and $p(\cdot)$ refer to transforming $\mathbf{m}_i^{crs}$ and $\mathbf{m}_i^{gen}$ into attention distributions through a softmax layer, and D_{KL} represents the KL divergence.

2.3 Model Training and Inference

During training, the visual representation $\mathbf{H}_v^{crs}$ and textual representation $\mathbf{H}_t^{crs}$ are first fed into a gated network (Xu et al., 2022) for fusion, and then connected with a CRF layer to computes the probability of the label, where the labeling loss is denoted as L_{crf}. During training, our method consists of two learning tasks, the CRF negtative log-likelihood loss for MNER task and the auxiliary loss for the AGA mechanism. We train the two tasks jointly,

and the final loss function is defined as follows:

$$L = L_{crf} + \lambda L_{align} \tag{11}$$

where λ is a hyper-parameter to control the contribution of the auxiliary loss.

During inference, we switch textual NER into MNER task by synthesizing image in Stage #1. The raw text and synthetic image are fed into Stage #2 for alignment and fusion. Then the obtained multimodal representations are input into CRF layer to predict the labels $Y = (y_1, y_2, \ldots, y_n)$.

3 Experiments

3.1 Datasets and Settings

We perform experiments on three representative datasets, Twitter-2015, Twitter-2017 and WikiDiverse (Wang et al., 2022), by excluding images that solely consist of textual corpus. Details of the datasets and metrics are described in the appendix A. We use the pre-trained stable diffusion 2.0-v to synthesize images[1]. During the image selection stage, we set the threshold θ to 0.05, η to 0.2 and the visual vocabulary size k to 100. For each textual sample, we generate 5 synthetic images and compare the similarity with 5 internet-crawled images, where $\alpha = 3$ denotes the number of selected images for the first filtering procedure. We set the maximum sentence length n to 128, the epochs to 25, the mini-batch size to 32, the hidden dimension d to 768, and the number of attention heads to 12. The Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of 5e-5, a dropout rate of 0.9, and an auxiliary loss weight λ of 0.5.

3.2 Baselines

3.2.1 Text-Only Models

(1) `BiLSTM-CRF` (Huang et al., 2015) is a classic NER model stacking a bidirectional LSTM layer and a CRF layer. (2) `BERT-CRF` (Liu et al., 2020) employs BERT as the encoder and CRF as the decoder. (3) `BERT+BS` (Zhu and Li, 2022) proposes boundary smoothing as a regularization technique for span-based NER models. (4) `MultiNER` (Wang et al., 2023b) proposes a multi-task learning framework for MRC-based NER.

3.2.2 Text-Image Models

(1) `UMT` (Yu et al., 2020) presents a multimodal interaction module to generate expressive text-visual representations. (2) `MT` (Yu et al., 2020) is the variation of UMT with the ablation of auxiliary entity span detection. (3) `RpBERT` (Sun et al., 2021) proposes a novel text-image relation propagation based multimodal BERT model. (4) `UMGF` (Zhang et al., 2021a) proposes a unified graph fusion approach to obtain the text-image representation. (5) `MAF` (Xu et al., 2022) is a general matching and alignment framework for the MNER task. (6) `CogVLM` (Weihan Wang, 2023) is the latest multimodal large model, fine-tuned here with LoRA (Hu et al., 2022) to serve as the newest baseline for comparison.

3.3 Main Results

The main results are reported in Table 1, from which we draw the following conclusions.

First, MNER methods generally perform better than text-only NER methods. Comparing the multimodal SOTA model `MAF` with its unimodal counterpart `MultiNER`, the former achieves F1 gains of 0.63%, 1.54%, and 1.11% on three datasets, respectively. The observation validates our motivation that image information functions as auxiliaries to provide enriched context for text-based NER. However, the fine-tuned CogVLM exhibits poor performance, attributed to the inherent limitations of generative multimodal large models in handling natural languange understanding tasks ($i.e.$ NER), for which comprehension based models like BERT are inherently more adept.

Second, solely relying on the textual corpus of three datasets, `InstructNER` achieves F1 scores of 74.12%, 86.28%, and 74.46%, surpassing all text-only models. Notably, our method outperforms the best text-only model, `MultiNER`, by 1.41%, 2.30%, and 2.13%, demonstrating the efficacy of synthetic images as a supplement for enhancing textual NER.

Third, comparing Twitter-2017 with Twitter-2015 and WikiDiverse, there exists a discrepancy of 12.16% and 11.82% of `InstructNER`'s F1-scores. The discrepancy is attributed to the shorter sentence length and the chaos derived from more ungrammatical sentences in Twitter-2015, like the internet slang *YOLO, man*. In WikiDiverse, the presence of multiple topics and more entity categories also increase the complexity of identification.

Last, when compared to existing multimodal methods that rely on sophisticated annotated images in datasets, our method still achieves remarkable results. `InstructNER` yields a substantial enhancement by 0.78%, 0.76%, and 1.02% compared to the best MNER method `MAF`. It is noteworthy that

[1]https://github.com/Stability-AI/stablediffusion

Table 1: Performance comparison (%) on three datasets, where the best results are in bold. All results are obtained by running the code released by the author. We report the average performance over 5 runs with random initialization, with InstructNER significantly better than MultiNER and MAF with p-value < 0.05 based on paired t-test.

Model	Twitter-2015			Twitter-2017			WikiDivere		
	P	R	F1	P	R	F1	P	R	F1
Text-Only Datasets									
BiLSTM-CRF (Huang et al., 2015)	68.14	61.09	64.42	79.42	73.43	76.31	67.32	60.19	63.55
BERT-CRF (Liu et al., 2020)	69.22	74.59	71.81	83.32	83.57	83.44	69.61	73.17	71.34
BERT+BS (Zhu and Li, 2022)	71.34	73.34	72.32	83.24	84.12	83.68	71.27	72.81	72. 03
MultiNER (Wang et al., 2023b)	71.42	74.05	72.71	83.82	84.14	83.98	71.25	73.45	72.33
Text & Raw Image Datasets									
CogVLM (Weihan Wang, 2023)	65.37	59.42	62.03	69.86	66.43	68.10	65.43	63.58	64.49
MT (Yu et al., 2020)	71.24	74.17	72.68	84.04	85.34	84.69	70.80	72.90	71.84
UMT (Yu et al., 2020)	71.84	74.61	73.20	85.08	85.27	85.18	71.75	73.77	72.75
RpBERT (Sun et al., 2021)	70.93	74.94	72.88	84.27	85.80	85.03	72.77	73.84	73.30
UMGF (Zhang et al., 2021a)	71.54	74.59	73.03	85.30	84.99	85.14	-	-	-
MAF (Xu et al., 2022)	71.75	**75.01**	73.34	85.39	85.65	85.52	73.03	73.86	73.44
Text & Synthetic Image Datasets									
InstructNER	**73.41**	74.84	**74.12**	**86.22**	**86.34**	**86.28**	**74.08**	**74.84**	**74.46**

our approach achieves the highest enhancement of 1.02% on WikiDiverse. This can potentially be attributed to the fact that WikiDiverse is sourced from the news domain, encompassing more enriched and diverse topics. Our method is capable of synthesizing images that incorporate various topic-specific scenarios. Furthermore, WikiDiverse includes more diverse entity categories, such as *APP* and *Film*. Through the guidance of the AGA mechanism, entities of these categories are more readily aligned with corresponding regions in images.

3.4 Ablation Experiments

To access individual component efficacy, we conduct ablation experiments on the full `InstructNER` model and its variants. As depicted in Table 2, we observe that the two image selection strategies (EC and EA) and the AGA mechanism significantly contribute to the final results. The performance decline follows the order: `InstructNER w/o AGA` >`InstructNER w/o EA` >`InstructNER w/o EC`. Specifically, removing the `EC` module results in a decrease of 0.49%, 0.43%, and 0.41% in the F1 score, while the removal of the `EA` module leads to performance drops of 0.87%, 0.82%, and 0.64%. This highlights the relatively more substantial factor by the `EA` than `EC` in improving performance. The primary reason behind this observation might be that the entity nouns are consistently portrayed in synthetic images, yet the fidelity or accuracy of these entities might not invariably be upheld. Moreover, removing the `AGA` module has the most significant decline of 1.17%, 1.20%, and 1.04%. This supports the critical importance of `AGA` technique in aligning text and images.

Table 2: Results of ablation experiments $w.r.t.$ the attention-guided alignment (AGA) module, two image selection strategies considering entity coverage (EC) and entity accuracy (EA). Here, ↓ represents the performance declines of variant models.

Models	Twitter-2015		Twitter-2017		WikiDiverse	
	F1(%)	↓(%)	F1(%)	↓(%)	F1(%)	↓(%)
InstructNER	**74.12**	-	**86.28**	-	**74.46**	-
w/o EC	73.63	0.49	85.85	0.43	74.05	0.41
w/o EA	73.25	0.87	85.46	0.82	73.82	0.64
w/o AGA	72.95	1.17	85.08	1.20	73.42	1.04

3.5 Analysis Experiments

3.5.1 Effect under Different Sample Numbers

To examine the robustness of our models under varying dataset sizes, we conduct experiments by randomly sampling 2000 to 500 instances from three datasets. As shown in Figure 3, the performance of `InstructNER` is compared against two baseline models, `MultiNER` and `BERT-CRF`. It can be observed that when the number of training samples decreases, the F1 scores of the two text-only baseline models both decline rapidly. However, the decline trend of InstructNER is relatively slow. This observation indicates that recognizing named entities based solely on textual context becomes increasingly challenging when the training sample size is limited. To address this point, incorporating images as supplementary information significantly enhances the robustness of NER.

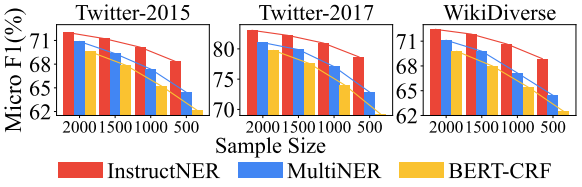

Figure 3: Performance of InstructNER, BERT-CRF and BiLSTM-CRF under different sample numbers.

3.5.2 Effect of Multi-Task Learning Losses

We examine and illustrate the influence of the auxiliary loss weight λ on the AGA mechanism in Figure 4, providing valuable insights into its impact. Our first observation reveals that as λ varies from 0.3 to 0.7, the F1 scores for both datasets exhibit an upward trend, reaching their peak around the value of 0.5. However, further increasing the value of λ leads to a decline in F1 scores. Essentially, a higher value of λ indicates a stronger emphasis on text-image alignment within the model. Conversely, as λ decreases, the model places greater importance on learning the entity classification. When the losses from both modules are approximately balanced at a value of 0.5, the model achieves its highest overall performance. This balance allows for effective integration of both text-image alignment and entity classification, resulting in superior results across the board.

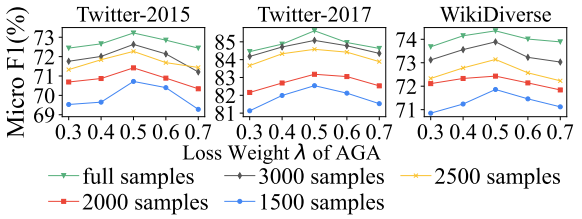

Figure 4: Performance of InstructNER under different multi-task learning weights on three datasets.

3.5.3 Effect of Synthetic Image Datasets

To validate the synthetic image datasets constructed by our method, we substitute the images in raw datasets with our synthetic ones and evaluate the performance on several existing MNER models. As shown in Table 4, our synthetic images exhibit comparable results to original images across all three models. Considering MAF model, using the synthetic images only results in a marginal decrease of 0.22%, 0.13%, and 0.18% in performance. This underscores the effectiveness of our synthetic image dataset. Without any manual annotation, our synthetic image datasets have achieved results closely resembling the original images at an remarkably low cost. Furthermore, even when using slightly less effective synthetic images, the inclusion of the AGA mechanism has allowed our method to surpass MAF with raw images. This demonstrates the robustness of our AGA mechanism.

Table 3: Results of different models on our synthetic image dataset (F1-score). Raw and Syn respectively represent original images and synthetic images.

Methods	Twitter-2015		Twitter-2017		WikiDiverse	
	Raw	Syn	Raw	Syn	Raw	Syn
UMT	73.20	73.07	85.18	85.03	72.75	72.65
RpBERT	72.86	72.74	85.03	84.81	73.30	73.57
MAF	73.34	73.12	85.52	85.39	73.44	73.26

3.5.4 Effect under Different Sentence Length

To validate that InstructNER provides more auxiliary information for text-only NER tasks under the challenging condition of short sentence lengths, we categorize samples in the test sets based on sentence length and separately record recognition results for the text-only methods and InstructNER in Table 4. For sentences shorter than 10 words, MultiNER achieves the lowest F1 score at 70.11%, 81.68% and 69.69% on three datasets, underscoring the limitation of relying solely on textual context, particularly in shorter sentences. However, in this scenario, InstructNER shows the highest improvement in F1 score compared to MultiNER, reaching 1.75%, 3.08% and 2.55%. This suggests that our synthetic images provide more supplemental information when the textual context is limited.

Table 4: Results (F1-score) across samples with varing sentence length. Here, L refers to sentence length.

Methods	Twitter2015			Twitter2017			WikiDiverse		
	L<10	L:10-20	L>20	L<10	L:10-20	L>20	L<10	L:10-20	L>20
BERT-CRF	69.31	72.46	72.54	80.96	84.25	84.37	69.11	72.14	72.25
MultiNER	70.11	73.28	73.41	81.68	84.78	84.86	69.69	73.07	73.13
InstructNER	71.86	74.77	74.83	84.76	86.83	86.85	72.24	75.18	75.24

3.5.5 Effect of BOW Strategy

We employ the siamese network approaches as an alternative to the BOW strategy, verifying the effect of BOW for image selection with entity accuracy. Specifically, pre-trained models VGG (Simonyan and Zisserman, 2015) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) are utilized to extract image features for similarity measurement. The selected synthetic images are then fed into our model to perform MNER. Results in Table 5 demonstrate thay the utilization of the BOW strategy in extraction region-level image features, coupled with the comparison of fine-grained similarities is overall superior to the siamese network

methods. Notably, the BOW strategy achieve score improvements of 0.51%, 0.13%, 0.29% compared to ViT encoder, supporting for our motivation in adopting the BOW strategy for the image selection.

Table 5: Results of Different Image Selection Methods.

Methods	Twitter-2015			Twitter-2017			WikiDiverse		
	P	R	F1	P	R	F1	P	R	F1
VGG	72.47	74.36	73.40	85.74	85.98	85.86	73.34	74.57	73.95
ViT	72.64	74.62	73.61	85.89	86.22	86.05	73.66	74.68	74.17
BOW	**73.41**	74.84	**74.12**	**86.22**	**86.34**	**86.28**	**74.08**	**74.84**	**74.46**

3.6 Case Study

Figure 5 illustrates two representative NER cases comparing InstructNER, MultiNER, and MAF. Regarding the first example, MultiNER incorrectly labels the named entity *Iman Shumpert* as O. However, with the aid of the image containing *Iman Shumpert*, MAF and InstructNER accurately align the text with the corresponding image region, and assign the correct labels. Regarding the second example, both MAF and MultiNER misclassify *Loch Lomond*, whereas InstructNER leverages attention guidance to focus more on the *Loch Lomond* lake, leading to the correct classification as *LOC*.

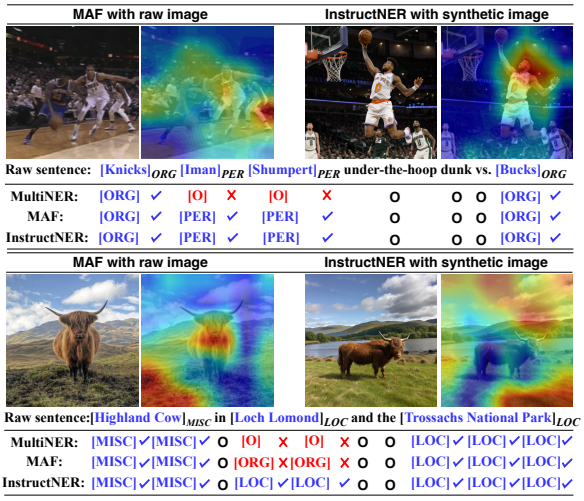

Figure 5: Two representative cases. For each case, the left two images correspond to MAF, while the right two images correspond to InstrucNER. The heat-maps correspond to *Iman Shumpert* and *Loch Lomond*.

4 Related Work

4.1 Textual NER and MNER

Traditional NER methods combine various neural network architectures with a CRF layer (Ratinov and Roth, 2009) to perform sequence labeling (Huang et al., 2015; Ma and Hovy, 2016; Yang et al., 2018). BERT (Devlin et al., 2019), a pre-trained language model, has also shown impressive results. To enhance NER, Zhu and Li (2022) propose boundary smoothing as a regularization technique for span-based NER. To address the issue of lacking context (Zhang et al., 2020; Ju et al., 2020) in text-only NER methods, MNER is introduced by utilizing image information as supplement. (Zhao et al., 2022; Wang et al., 2023a). However, obtaining high-quality images is costly. In this study, we propose utilizing stable diffusion models for image synthesis to provide additional information.

4.2 Text-Image Alignment

Existing text-image alignment approaches can be categorized into explicit and implicit alignment methods. Explicit methods involve extracting object regions from images and aligning them with corresponding words. Zhang et al. (2021b) utilize a visual grounding toolkit (Yang et al., 2019b) to ground sentences to image regions. Jia et al. (2022) further design queries of label types to enhance the association between regions and tokens. However, explicit alignment methods may suffer from error propagation with inaccurate object regions (Yang et al., 2019b). Implicit methods address this issue by using attention mechanisms to adaptively align texts and images. Yu et al. (2020) introduces a multimodal interaction module that integrates transformer layers with cross-modal attention to perform a hierarchical alignment. To eliminate noise, Zong and Sun (2023) aggregates visual features into bottleneck tokens and propagates the refined tokens into alignment. Despite the efficiency of these approaches, attention may lack concentration and contain irrelevant noise. Therefore, we utilize the by-product cross-attention matrix of stable diffusion to guide soft alignment and facilitate attention concentration.

5 Conclusion

In this paper, we propose InstructNER, a comprehensive method that utilizes the image synthesis capability of the stable diffusion model, in order to reframe the textual NER as the MNER task and learn instructive alignment between synthetic images and raw texts. The experimental results demonstrate the effectiveness of synthetic images and the innovative soft attention guidance alignment (AGA) mechanism in improving NER performance. Moreover, we have made available a large-scale, high-quality dataset of synthetic images, which complements existing raw datasets and provides valuable insights for future research endeavors.

Limitations

Our method employs the image synthesis capability of stable diffusion to provide additional context for text-only NER tasks, achieving significant results in general domains. However, domain-specific image synthesis methods, like the medical and electrical domains, encounter challenges in transforming text into images (Kazerouni et al., 2023), resulting in suboptimal outcomes. Regrettably, there is limited current research addressing domain-specific MNER tasks. Therefore, our approach may not be suitable for specialized domains. Furthermore, recent studies have proposed pre-training stable diffusion in specific domains (Moghadam et al., 2023; Kazerouni et al., 2023), which can partially mitigate this issue.

References

Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal named entity recognition with image attributes and image knowledge. In *DASFAA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. 2020. Learning representations by predicting bags of visual words. In *CVPR*.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. In *ICRL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *ICLR*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang, Lejian Liao, Yang Song, Meng Chen, and Xiaodong He. 2022. Query prior matters: A mrc framework for multimodal named entity recognition. In *ACM MM*.

Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *ACM MM*.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. 2023. Diffusion models for medical image analysis: A comprehensive survey. In *Medical Image Analysis*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *ACL*.

Mingyi Liu, Zhiying Tu, Zhongjie Wang, and Xiaofei Xu. 2020. Ltp: a new active learning strategy for bert-crf based named entity recognition. *arXiv preprint arXiv:2001.02524*.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *ACL*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *ACL*.

P. Moghadam, S. Van Dalen, K. C. Martin, J. Lennerz, S. Yip, H. Farahani, and A. Bashashati. 2023. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *WACV*. IEEE Computer Society.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *EACL*. ACL.

K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. In *AAAI*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Peng Wang, Xiaohang Chen, Ziyu Shang, and Wenjun Ke. 2023a. Multimodal named entity recognition with bottleneck fusion and contrastive learning. *IEICE Transactions*, 106(4):545–555.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *ACL*.

Yibo Wang, Wenting Zhao, Yao Wan, Zhongfen Deng, and Philip S. Yu. 2023b. Named entity recognition via machine reading comprehension: A multi-task learning approach.

Wenmeng Yu Wenyi Hong Ji Qi Yan Wang Junhui Ji Zhuoyi Yang Lei Zhao Xixuan Song Jiazheng Xu Bin Xu Juanzi Li Yuxiao Dong Ming Ding Jie Tang Weihan Wang, Qingsong Lv. 2023. Cogvlm: Visual expert for pretrained language models.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *ACM MM*.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: A general matching and alignment framework for multimodal named entity recognition. In *WSDM*.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *COLING*.

Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo. 2019a. A fast and accurate one-stage approach to visual grounding. In *ICCV*.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019b. A fast and accurate one-stage approach to visual grounding. In *ICCV*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL*.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multimodal multi-label emotion detection with modality and label dependence. In *EMNLP*.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *AAAI*.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021b. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *AAAI*.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweetsw. In *AAAI*.

Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *ACM MM*.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532.

Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *EMNLP*.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *ACL*.

Daoming Zong and Shiliang Sun. 2023. Mcomet: Multimodal fusion transformer for physical audiovisual commonsense reasoning. In *AAAI*.

A Datasets Details

We perform experiments on three representative English datasets, Twitter-2015, Twitter-2017 and WikiDiverse, by excluding images that solely consist of textual corpus. Note that WikiDiverse is a cutting-edge multimodal entity linking dataset based on WikiNews. We transform WikiDiverse into a MNER dataset to further verify the effectiveness on the news domain. The statistic are shown in Table 6. Among these, Twitter2015 and Twitter2017 encompass four types of entity categories: Person, Organization, Location, and Other. WikiDiverse contains seven categories, namely: Person, Organization, Country, Movie, Event, Building, and Other. We utilize precision (P), recall (R), and micro F1 score (F1) to evaluate the performance of named entity recognition for all datasets.

Table 6: The statistics of the three MNER datasets.

Dataset	Domain	Types	Train	Dev	Test
Twitter-2015	Social Media	4	4000	1000	3257
Twitter-2017	Social Media	4	3373	723	723
WikiDiverse	News	7	6377	796	796