Decompose, Prioritize, and Eliminate: Dynamically Integrating Diverse Representations for Multi-modal Named Entity Recognition

Zihao Zheng[1], **Zihan Zhang**[2], **Zexin Wang**[1], **Ruiji Fu**[3]
Ming Liu[1,4], **Zhongyuan Wang**[3] **and Bing Qin**[1,4]
[1]Harbin Institute of Technology [2] Imperial College London
[3]Kuaishou Technology [4] Peng Cheng Laboratory
{zhzheng, mliu, qinb}@ir.hit.edu.cn

Abstract

Multi-modal Named Entity Recognition, a fundamental task for multi-modal knowledge graph construction, requires integrating multi-modal information to extract named entities from text. Previous research has explored the integration of multi-modal representations at different granularities. However, they struggle to integrate all these multi-modal representations to provide rich contextual information to improve multi-modal named entity recognition. In this paper, we propose DPE-MNER, which is an iterative reasoning framework that dynamically incorporates all the diverse multi-modal representations following the strategy of “decompose, prioritize, and eliminate”. Within the framework, the fusion of diverse multi-modal representations is **decomposed** into hierarchically connected fusion layers that are easier to handle. The incorporation of multi-modal information **prioritizes** transitioning from "easy-to-hard" and "coarse-to-fine". The explicit modeling of cross-modal relevance **eliminate** the irrelevances that will mislead the MNER prediction. Extensive experiments on two public datasets have demonstrated the effectiveness of our approach.

Keywords: Named Entity Recognition, Multi-modal Fusion, Iterative Reasoning

1. Introduction

In recent years, the emergence of multi-modal data on the Web, combining textual information with other modalities such as images, has added a new dimension to Named Entity Recognition (NER), giving rise to Multi-modal Named Entity Recognition (MNER) (Zhang et al., 2018; Lu et al., 2018). The goal of MNER is not only to use text to extract named entities but also to borrow contextual information from additional visual cues, thereby increasing the depth and breadth of information extraction (Yu et al., 2020; Jia et al., 2023).

MNER research focuses primarily on the fusion of multi-modal representations to improve NER (Zhang et al., 2018). So far, researchers have explored several multi-modal representations. For text representations, most work adopts the token-level representation (Yu et al., 2020). For image representations, there are coarse-grained Yu et al. (2020) and fine-grained image representations (Chen et al., 2022c), and there are also aligned image representations that are obtained by translating image modality to text (Wang et al., 2021)). However, none of these works make use of all these different multi-modal representations. For the text modality, there are sentence- (Gao et al., 2021), span- (Zhao et al., 2022), and token-level representations that capture text semantics at different scales; for the visual modality, currently, popular image representations can be categorized into coarse-grained (Yu et al., 2020) and fine-grained approaches (Zhang et al., 2021a). Under each granularity of image representation, they can be further categorized into representations obtained directly from visual foundation models (Dosovitskiy et al., 2020) as well as those obtained by translating the image into text space (Yang et al., 2022; Wang et al., 2021). Referring to the example in Fig. 1, all of these multi-modal representations should be synthesized for the decision-making of NER.

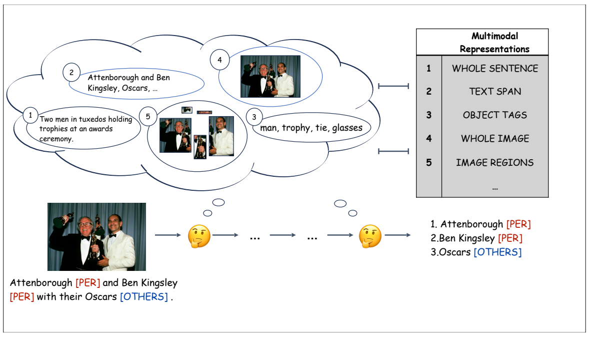
2
Attenborough and Ben Kingsley, Oscars, ...
4
1
Two men in tuxedos holding trophies at an awards ceremony.
5
3
man, trophy, tie, glasses
Multimodal Representations
1 WHOLE SENTENCE
2 TEXT SPAN
3 OBJECT TAGS
4 WHOLE IMAGE
5 IMAGE REGIONS
...
...
...
1. Attenborough [PER]
2.Ben Kingsley [PER]
3.Oscars [OTHERS]
Attenborough [PER] and Ben Kingsley [PER] with their Oscars [OTHERS] .

Figure 1: An example of multi-modal named entity recognition. We show various multi-modal representations that could be useful for NER decision-making. Humans usually process them iteratively in mind.

It is non-trivial to integrate all these different multi-modal representations (referred to as “diverse-modal multi-modal fusion” for simplicity in the rest of this paper) since there will inevitably be noisy data and the overwhelming dominance of one particular piece of information (Chen et al., 2022a). The fusion of these representations is a complex

LREC-COLING 2024, pages 4498–4508

problem that should be solved with specific strategies and procedures. To solve this problem, we are inspired by a field called “Complex Problem Solving” (Sternberg and Frensch, 1992), which studies the methods and strategies humans and computers use to solve problems involving multiple variables, uncertainty, and high complexity. When faced with complex problems, humans typically process them iteratively and employ certain strategies to simplify the complex problem, such as decomposition, prioritization, and elimination of irrelevance.

We argue that modeling MNER as an iterative process that integrates multi-modal information with these strategies is well-suited for MNER. Compared to single-step methods, multi-step approaches can more fully exploit the diverse multi-modal representations during the iterative refinement of NER results. Moreover, these three strategies are well suited for integrating multiple representations in multi-modal NER: 1) Decomposition inspires us to break down the fusion of diverse multimodal representations into smaller and easier fusion units, which explore multi-modal interactions at different granularities (Majumder et al., 2018). 2) Prioritization suggests integrating multi-modal information according to the “easy-to-hard” and “coarse-to-fine” priorities; this gradual integration helps the progressive refinement of MNER predictions. This allows the model to gradually shift its focus from simple but coarse information to challenging yet precise details. 3) Irrelevance elimination inspires us to explicitly filter out the irrelevant information in the different multi-modal representations; it can eliminate the irrelevant information that will hurt MNER performance (Sun et al., 2021).

Specifically, we devise a novel framework, DPE-MNER, which formulates MNER as an iterative process with an adaptive multi-modal reasoning network. For the design of multi-modal fusion, we follow the strategies of “decompose, prioritize, and eliminate” discussed earlier. To deploy the **decompose** strategy, we design hierarchically decomposed multi-modal fusion which decomposes complex multi-modal fusion into three layers of multi-modal fusion units: at the bottom layer, we fuse text representations of a certain granularity with image representations of the same granularity but with different gaps, resulting in multi-modal text representations conditioned on a certain granularity of the image representation; at the middle layer, we fuse multi-modal text representations conditioned on either coarse- or fine-grained image representations, generating the multi-modal representations of a specific text granularity; at the top layer, we fuse the multi-modal representations of different text granularities. To implement the **prioritize** strategy, during the iterative reasoning process, we propose the prioritized multi-modal information integration, which gradually integrates multi-modal information with "easy-to-hard" and "coarse-to-fine" prioritization in chronological order of the iterative reasoning steps, allowing the model to take full advantage of multi-modal information in multiple steps. To deploy the **eliminate** strategy, we develop explicit cross-modal relevance modeling in each layer of multi-modal fusion to eliminate the noises.

In summary, we have the following contributions:

- We are the first to model MNER as an iterative reasoning process that dynamically fuses diverse multi-modal representations, better aligning with the human decision process.
- We try to solve the difficulties in diverse multi-modal fusion by the “decompose, prioritize, and eliminate” strategies and design an adaptive multi-modal reasoning network with hierarchically decomposed multi-modal fusion, prioritized multi-modal information integration, and explicit cross-modal relevance modeling at multiple granularities.
- Experimental results demonstrate that our proposed DPE-MNER achieves new state-of-the-art results on two well-known MNER datasets.

2. Methodology

We introduce DPE-MNER in this section. We first give the Preliminaries (2.1) and then describe the Acquisition of Diverse Multi-modal Representations (2.2) and the Adaptive Multi-modal Reasoning Network (2.3) that can integrate diverse multi-modal representations to decode the entities in a new reasoning iteration. Figure 2 illustrates the overall framework.

2.1. Preliminaries

Task definition Given a multi-modal context $\{S, I\}$ where S is a sentence with length M and I is an image. The multi-modal named entity recognition task is to extract the entities $E = \{(span_i, type_i)\}_{i=0}^{N_s}$ contained in S with the help of I. N_s is the number of entities, and $span_i$ and $type_i$ are the span and type of a certain entity.

Modeling MNER as iterative reasoning process Inspired by the works that model object detection and grounding in images as an iterative reasoning process (Chen et al., 2022b; Chen and Li, 2023), we model MNER as an iterative reasoning process by denoising diffusion modeling to explore multi-modal information in multiple iterations extensively.

Diffusion models construct a Markov chain through a gradual introduction of noise to the initial data sample z_0, thus delineating the forward

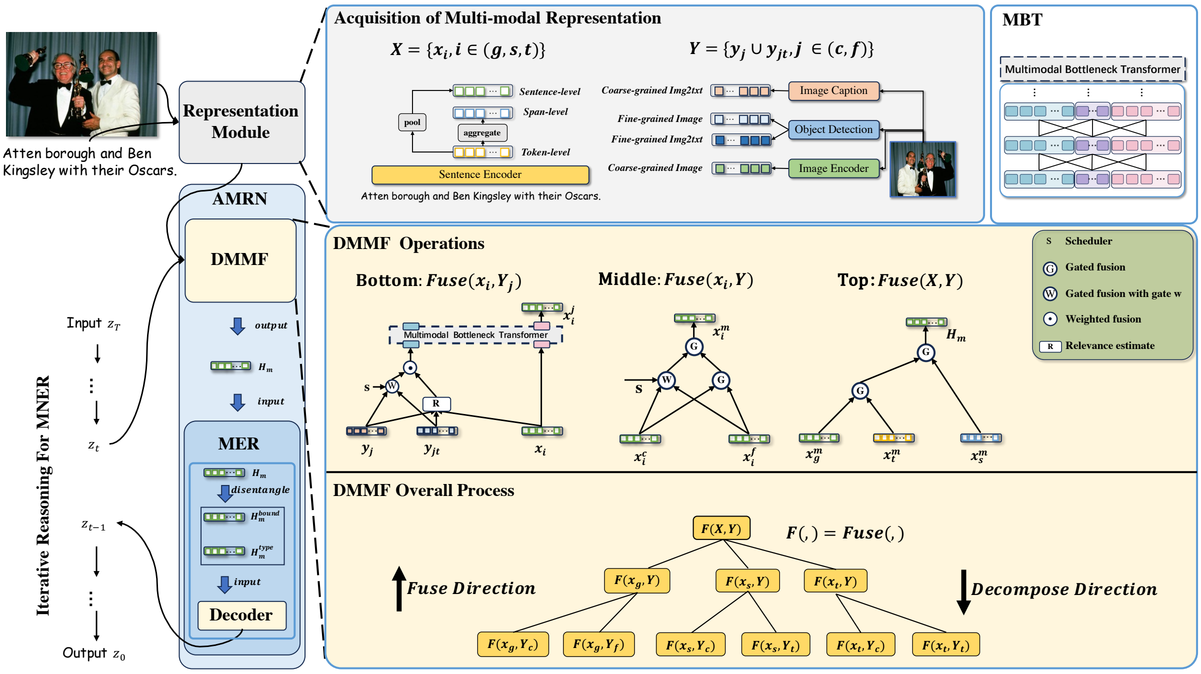

Figure 2: Overview of DPE-MNER. DPE-MNER models MNER as an iterative reasoning process in multiple steps. During this process, AMRN (Adaptive Multi-modal Reasoning Network) reasons the output in the next step. AMRN contains DMMF (Diverse Multi-modal Fusion) and MER (Multi-modal-guided Entity Reasoning) for encoding multi-modal information and decoding named entities. On the right are the key details about DMMF(We use $\boldsymbol{X}$ and $\boldsymbol{Y}$ to denote the different text and image representations to keep the figure clear). This framework is designed based on the "Decompose, Prioritize, and Eliminate" strategy.

diffusion process (Ho et al., 2020). This forward process is:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t|\sqrt{\alpha_t}\boldsymbol{z}_0, (1-\alpha_t)\boldsymbol{I}), \qquad (1)$$

where $\alpha_t := \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)$ with β_s denoting the noise variance schedule. During the training stage, a neural network $f_\theta(\boldsymbol{z}_t, t)$ is trained to predict $\boldsymbol{z}_0$ from $\boldsymbol{z}_t$, utilizing an ℓ_2 loss (Ho et al., 2020):

$$\mathcal{L}_{\text{train}} = \frac{1}{2}||f_\theta(\boldsymbol{z}_t, t) - \boldsymbol{z}_0||^2. \qquad (2)$$

During the inference stage, $\boldsymbol{z}_0$ is iteratively reconstructed from the noise $\boldsymbol{z}_T$ using the model f_θ (Ho et al., 2020).

To adapt this paradigm to MNER, we are inspired by Shen et al. (2023) and use entity span boundaries in the text as data samples $\boldsymbol{z}_0 = \boldsymbol{b}$, where $\boldsymbol{b} \in \mathbb{R}^{N_s \times 2}$ is a set of N_s entity span boundaries. A multi-modal reasoning network $f_\theta(\boldsymbol{z}_t, t, \boldsymbol{Multi{-}modal})$ is trained to predict $\boldsymbol{z}_0$ from noisy spans $\boldsymbol{z}_T$, conditioned on the multi-modal context. Then, the entity labels are correspondingly predicted (Shen et al., 2023; Chen et al., 2022b).

2.2. Acquisition of Diverse Multi-modal Representations

2.2.1. Text Representation

We obtain text representations at three different granularities to capture text semantics at different levels (Liu et al., 2023). For the conciseness of the subsequent fusion section, we use $\boldsymbol{X}$ to represent all these text representations, including token-level $\boldsymbol{x}_t$, sentence-level $\boldsymbol{x}_g$, and span-level $\boldsymbol{x}_s$.

- **Token-level text representation** We use BERT (Devlin et al., 2019) to encode the input sentence S and get the token-level text encoding $\mathbf{x}_t \in \mathbb{R}^{M \times h}$, where M is the sequence length, and h is the dimension of each token.
- **Sentence-level text representation** Following SimCSE (Gao et al., 2021), we use the encoding of [CLS] in $\mathbf{x}_t$ as the sentence-level text representation, which can capture the whole sentence information $\mathbf{x}_g \in \mathbb{R}^h$.
- **Span-level text representation** After we get $\mathbf{x}_t$, we aggregate it to span representations to capture span-level text semantics.

  In the iterative reasoning process, we can obtain the previously predicted entity spans.

Based on these spans, $\mathbf{x}_t$ is first mean-pooled with the span indexes to raw span-level representations $\mathbf{x}_{s0} \in \mathbb{R}^{N_s \times h}$. Then a span aggregator (SpanAggr) that consists of a self-attention and a cross-attention layer is used to further encode and aggregate the context from $\mathbf{x}_t$, and a position embedding $\mathbf{E}_t$ is added (Shen et al., 2023). Finally, we get the span-level text representations $\mathbf{x}_s \in \mathbb{R}^{N_s \times h}$.

2.2.2. Image Representation

We use four different representations covering coarse- and fine-grained image information. Here, we regard representations capturing entire image information as coarse-grained and those capturing fine-grained image information such as regions and objects as fine-grained. Furthermore, as stated in ITA (Wang et al., 2021), translating the image to text directly can reduce the modality gap, which benefits multi-modal interactions (which we call the “image2text” series). For conciseness, we use $\boldsymbol{Y}$ to represent all these text representations, including coarse-grained $\boldsymbol{Y}_c$ and fine-grained representations $\boldsymbol{Y}_f$. Under each representation, there are image $\boldsymbol{y}_j$ and image2text $\boldsymbol{y}_{jt}$ representations that have different modality gaps with text.

- **Coarse-grained image representation** To get this representation, we feed the image to ResNet to get the full image representation $\mathbf{y}_c \in \mathbb{R}^{2048 \times M_g}$, M_g is the number of visual blocks, and 2048 is the dimension of each block’s representation.

- **Coarse-grained image2text representation** We use VinVL (Zhang et al., 2021b) large model fine-tuned on image caption datasets to get the caption of the image. Then we use BERT-base to get the caption’s representation $\mathbf{y}_{ct} \in \mathbb{R}^{d \times N_g}$, d is the dimension of each token and N_g is the caption’s length.

- **Fine-grained image representation** We follow HVPNeT (Chen et al., 2022c) to get the regions. These regions are then rescaled and fed into ResNet to get the fine-grained image representation $\mathbf{y}_f \in \mathbb{R}^{2048 \times 3 \times M_l}$, and M_l is the number of regions in an image.

- **Fine-grained image2text representation** Following ITA (Wang et al., 2021), we use the object detection module of VinVL to detect the top 5 objects in the image. Then, the tags of these objects are concatenated and encoded by the BERT base. The fine-grained image2text representation is denoted as $\mathbf{y}_{ft} \in \mathbb{R}^{d \times N_l}$, N_l is the length of the concatenated tag sequence.

2.3. Adaptive Multi-modal Reasoning Network

Based on the reasoning process stated in Section 2.1, we propose an Adaptive Multi-modal Reasoning Network (abbreviated as AMRN), which accepts the previous entity predictions and reasons the predictions in the next iteration step.

Concretely, AMRN contains two sub-modules: Diverse Multi-modal Fusion and Multi-modal-guided Entity Reasoning. Diverse Multi-modal Fusion fuses multi-modal information in a dynamic way. Multi-modal-guided Entity Reasoning infers the entities for the current iteration step based on the fused multi-modal information.

2.3.1. Diverse Multi-modal Fusion

Diverse Multi-modal Fusion has a three-layered hierarchical structure based on the "Decompose, Prioritize, and Eliminate" strategies. These three strategies lead to three key designs. **Decompose** results in three-layer hierarchically decomposed multi-modal fusion, comprising three multi-modal fusion operations: bottom-layer ($Fuse(\boldsymbol{x}_i, \boldsymbol{Y}_j)$), middle-layer ($Fuse(\boldsymbol{x}_i, \boldsymbol{Y})$), and top-layer ($Fuse(\boldsymbol{X}, \boldsymbol{Y})$) fusions. **Prioritize** leads to prioritized multi-modal information integration, which gradually integrates multi-modal information with "easy-to-hard" and "coarse-to-fine" prioritization in chronological order information in multiple steps during the iterative reasoning process. **Eliminate** leads to the explicit cross-modal relevance modeling, which occurs in each layer of multi-modal fusion to eliminate the noises.

Prioritized multi-modal information integration and explicit cross-modal relevance modeling are coupled tightly into these operations for cooperation. Next, we will detail these three layers from bottom to top.

Bottom-layer multi-modal fusion In this layer, we combine text representations and image representations at a specific granularity. There are two types of image representations at the same granularity but have different degrees of modality gap with text. The fusion of the text representation and image2text representations is usually regarded as easy since the model does not need to overcome the modality gap. In contrast, the fusion between the text and image representations is deemed hard. However, the image2text representations will lose information during the cross-modal translation. To this end, we deploy the “prioritize” strategy to fuse text representation with image representations at different modality gaps with “easy-to-hard” priority dynamically. We employ the schedule to control the ratio of these two representations to make the "easy-to-hard" integration smoother. We have:

$$\mathbf{H}_{vj} = [(1-\lambda_x \cdot s) \odot \mathbf{y}_j || \lambda_x \cdot s \odot \mathbf{y}_{jt}]. \tag{3}$$

where λ_x is a hyperparameter that controls the minimum proportion of each representation. For s, we use $\sqrt{\alpha_t}$. It cannot be ignored that text can be irrelevant to the image; therefore, we deploy the "eliminate" strategy and add explicit cross-modal relevance here. We calculate a relevance gate rel^i to estimate the relevance between $\mathbf{x}_i$ and $\mathbf{y}_j$.

$$rel^i = \mathrm{Sigmoid}\left(\mathrm{FC}\left[\mathbf{x}_i; \mathbf{y}_j; \mathbf{y}_{jt}\right]\right) \tag{4}$$

where FC is the fully-connected layer. Then, this gate is used to calculate the coarse-grained contextual text representations $\mathbf{x}_i^j$ as follows:

$$\mathbf{x}_i^j = \mathrm{MBT}(\mathbf{x}_i, rel^i \cdot \mathbf{H}_{vj}) \tag{5}$$

where $i \in \{t, g, s\}$, and $j \in \{c, f\}$. MBT is the bottleneck Transformer proposed by Nagrani et al. (2021).

Middle-layer multi-modal fusion After we get the coarse- and fine-grained contextual text representations, we fuse them with and without respect to the time dynamics to get two kinds of text-conditioned multi-modal representations since the proportions of these two representations are determined by the dynamic reasoning phase as well as the inherent requirements of the samples. The former fusion is the second deployment of the "prioritize" strategy with "coarse-to-fine" priority smoothly. During the iterative process, it is a natural design to gradually incorporate multi-modal information in a "coarse-to-fine" manner to focus on the fine-grained details of input data.

We use the same scheduler stated in bottom-layer fusion to make the "coarse-to-fine" integration smoother:

$$\mathbf{x}_i^{ms} = \lambda_y \cdot s \odot \mathbf{x}_i^c + (1 - \lambda_y \cdot s) \odot \mathbf{x}_i^f \tag{6}$$

where λ_y is a hyperparameter used to control the minimum of each representation. We also fuse these two representations without using the scheduler with a gated fusion strategy.

$$\mathbf{x}_i^{mg} = g^{tr} \odot \mathbf{x}_i^c + (1 - g^{tr}) \odot \mathbf{x}_i^f \tag{7}$$

$$g^{tr} = \sigma([\mathbf{x}_i^c; \mathbf{x}_i^f]\mathbf{W}^1 + \mathbf{b}^1) \tag{8}$$

where $\mathbf{W}^1$ and $\mathbf{b}^1$ are trainable weights. Then $\mathbf{x}_i^{ms}$ and $\mathbf{x}_i^{mg}$ are fused as follows:

$$\mathbf{x}_i^m = g^{mx} \odot \mathbf{x}_i^{ms} + (1 - g^{mx}) \odot \mathbf{x}_i^{mg} \tag{9}$$

$$g^{mx} = \sigma([\mathbf{x}_i^{ms}; \mathbf{x}_i^{mg}]\mathbf{W}^2 + \mathbf{b}^2) \tag{10}$$

where $\mathbf{W}^2$ and $\mathbf{b}^2$ are trainable weights.

Top-layer multi-modal fusion We use sentence-level, span-level, and token-level text representations to get multi-modal representations conditioned on different text granularities.

We can obtain multi-modal representations that capture interactions at span-level $\mathbf{x}_s^m$, token-level $\mathbf{x}_t^m$, and sentence-level $\mathbf{x}_g^m$. To fuse these representations, we first aggregate the token-level representations and sentence-level representations to span representations $\mathbf{x}_{t2s}^m$ and $\mathbf{x}_{g2s}^m$.

$$\mathbf{x}_{t2s}^m = \mathrm{SpanAggr}(\mathbf{x}_s^m, \mathbf{x}_t^m) + \mathbf{E}_t \tag{11}$$

$$\mathbf{x}_{g2s}^m = \mathrm{SpanAggr}(\mathbf{x}_s^m, \mathbf{x}_g^m) + \mathbf{E}_t \tag{12}$$

Then, these representations are fused with a hierarchical two-stage gated fusion strategy:

$$\mathbf{H}_{x2s} = g^{x2s} \odot \mathbf{x}_{t2s}^m + (1 - g^{x2s}) \odot \mathbf{x}_{g2s}^m \tag{13}$$

$$g^{x2s} = \sigma([\mathbf{x}_{t2s}^m; \mathbf{x}_{g2s}^m]\mathbf{W}^3 + \mathbf{b}^3) \tag{14}$$

$$\mathbf{H}_m = g^{tx} \odot \mathbf{H}_m^{x2s} + (1 - g^{tx}) \odot \mathbf{x}_s^m \tag{15}$$

$$g^{tx} = \sigma([\mathbf{H}_m^{x2s}; \mathbf{x}_s^m]\mathbf{W}^4 + \mathbf{b}^4) \tag{16}$$

where $\mathbf{W}^3, \mathbf{W}^4$ and $\mathbf{b}^3, \mathbf{b}^4$ are trainable weights. Finally, $\mathbf{x}^m$ is fed into the Entity Decoder, explained in the following part.

2.3.2. Multi-modal-guided Entity Reasoning

We use the contextualized span representations to decode the locations and types.

In contrast to previous approaches that regard the multi-modal information in localization and classification as equally important (Chen et al., 2022c; Lu et al., 2022), we feed $\mathbf{H}_m$ to two separate fully connected layers to disentangle the multi-modal span representations into boundary-focused representations $\mathbf{H}_m^{bound}$ and type-focused representations $\mathbf{H}_m^{type}$.

Then we use the multi-modal representations to predict the boundaries and entity types (Shen et al., 2023). We use two boundary pointers to predict the entity boundaries. For boundary $\delta \in \{l, r\}$, we compute the fusion representation $\mathbf{H}_{DEC}^\delta \in \mathbb{R}^{N_s \times M \times h}$ of the noisy spans and the words, and compute the probability of the word as the left or right boundaries $\mathbf{P}^\delta \in \mathbb{R}^{N_s \times M}$ as follows:

$$\mathbf{H}_{DEC}^\delta = \mathbf{H}_m^{bound}\mathbf{W}_m^\delta + \mathbf{H}_t\mathbf{W}_t^\delta$$
$$\mathbf{P}^\delta = \mathrm{sigmoid}(\mathrm{FC}(\mathbf{H}_{DEC}^\delta))$$

where $\mathbf{W}_m^\delta, \mathbf{W}_t^\delta$ are trainable weights. The calculated boundary probabilities can be used to decode the boundary indices of the N_s noisy spans.

Furthermore, the classification probability $\mathbf{P}^c \in \mathbb{R}^{N_s \times \mathcal{C}}$ of the noisy spans is calculated as follows:

$$\mathbf{P}^c = \mathrm{softmax}(\mathrm{FC}(\mathbf{H}_m^{type}))$$

where $\mathcal{C}$ is the number of entity types.

2.4. Training Objectives

We have two training objectives: NER loss and cross-modal relevance loss.

NER loss uses the Hungarian algorithm to find the optimal matching π between the golden and predicted entity sets (Zhu et al., 2020; Shen et al., 2023). $\pi(i)$ denotes the entity which corresponds to i-th span. The NER loss is as follows:

$$\mathcal{L}_{NER} = -\sum_{i=1}^{N_s} \sum_{\delta \in \{l,r,c\}} \log \mathbf{P}_i^{\delta}\left(\pi^{\delta}(i)\right)$$

Cross-modal relevance loss is used to learn the relevance between the representations of the image and the text at different granularities. We use a self-supervised contrastive loss inspired by MAF (Xu et al., 2022), where we maximize the relevances between image and text belonging to the same image-text pair and minimize those belonging to different image-text pairs in the batch. This can facilitate the learning of image-text relevance and force the model to generate visually grounded entities. We use $rel^i(a,b)$ to denote the relevance between the text and image from the a-th and b-th instance in the mini-batch, where $i \in \{t,s,g\}$. The relevance loss is as follows:

$$\mathcal{L}_a^{T2I,i} = -\log \frac{\exp\left(rel^i(a,a)/\tau\right)}{\sum_{b=1}^{batch} \exp\left(rel^i(a,b)\right)/\tau)} \quad (17)$$

$$\mathcal{L}_b^{I2T,i} = -\log \frac{\exp\left(rel^i(b,b)/\tau\right)}{\sum_{a=1}^{batch} \exp\left(rel^i(a,b)\right)/\tau)} \quad (18)$$

$$\mathcal{L}_{align}^{i} = \frac{1}{N}\sum_{a=1}^{N}(\lambda_{rel}\mathcal{L}_a^{T2I,i} + (1-\lambda_{rel})\mathcal{L}_b^{I2T,i}) \quad (19)$$

Combining the NER loss and relevance losses, we train the tasks jointly:

$$\mathcal{L} = \mathcal{L}_{NER} + \mathcal{L}_{align}^{t} + \mathcal{L}_{align}^{s} + \mathcal{L}_{align}^{g} \quad (20)$$

3. Experiments

We use two publicly available Twitter datasets (Twitter2015 and Twitter2017), which are provided by (Zhang et al., 2018) and (Lu et al., 2018), respectively. We use the same metrics as in these works (Zhang et al., 2018).

3.1. Baselines

We compare our approach against three distinct categories of baseline methods.

The first category represents a collection of text-based NER methods including: (1) *BiLSTM-CRF* (Lample et al., 2016): A biLSTM followed by a CRF decoder, (2) *BERT-CRF*: A multilayer bidirectional transformer encoder followed by a CRF decoder, (3) *BERT-Span* (Wang et al., 2022): A span-based NER model with BERT as the backbone, (4) *BERT-Iterative* (Shen et al., 2023): A diffusion-based iterative NER framework utilizing BERT for reasoning.

The second category encompasses the popular multi-modal NER methods, including (1) *Ada-CoAtt* (Zhang et al., 2018): A CNN-BiLSTM-CRF equipped with an adaptive co-attention network to induce visual information, (2) *UMT* (Yu et al., 2020): A Transformer-based multi-modal interaction module with an auxiliary entity span detection module, (3) *CAT-MNER* (Wang et al., 2022): A span-based MNER model with refined multi-modal interaction, (4) *FMIT* (Lu et al., 2022): A method using unified lattice structure and entity boundary detection for fine-grained multimodal interaction, (5) *HVPNet* (Chen et al., 2022c): A hierarchical visual prefix to multi-layered multimodal interaction with pretrained language model, (6) *MNER-QG* (Jia et al., 2023): An MRC-based method that jointly learns visual grounding and named entity recognition, (7) *DebiasCL* (Zhang et al., 2023): A method that implicitly aligns text and image representations through debias contrastive learning. (8) *VisualPT-MoE* (Xu et al.): A method that incorporates all the diverse image representations at once with MoE to help MNER.

We choose two representative LLM baselines for the third category, ChatGPT and GPT-4. The results are taken from an existing paper (Chen and Feng, 2023). Please refer to the original paper for more details on the prompt strategies.

3.2. Implementation Details

The implementation details of our NLP research are as follows. We leverage an NVIDIA A100 GPU for computational efficiency. The choice of pre-trained model is BERT-base-cased (Devlin et al., 2019), which we fine-tune for our specific task. Our experiments are conducted over 100 epochs, with a batch size of 64. We employ the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2e-5. During iterations, we set the timestep to 1000 steps, and the number of noisy spans is 150. Hyperparameters λ_x and λ_y are set to 0.3, while λ_{rel} is chosen as 0.5. We utilize PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) for implementation.

3.3. Main Results

(1) From Table 1, we show the overall results of our model and the compared systems on Twitter15 and Twitter17. The table is divided into three parts, and from the comparison of LSTM and BERT, we can draw several important conclusions. First, using a

Methods	Twitter2015							Twitter2017						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
Text Baselines														
BiLSTM-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	78.16	80.37
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
BERT-Span	85.35	81.88	62.06	43.23	75.52	73.83	74.76	90.84	85.55	81.99	69.77	85.68	84.60	85.14
BERT-Iterative	86.64	83.58	63.15	44.35	75.42	76.76	76.09	91.81	82.95	85.28	68.84	87.82	84.83	86.30
LLM Baselines														
ChatGPT	-	-	-	-	-	-	50.21	-	-	-	-	-	-	57.5
GPT-4	-	-	-	-	-	-	57.98	-	-	-	-	-	-	66.61
Multi-modal Baselines														
AdaCoAtt	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
CAT-MNER	85.57	82.53	63.77	43.38	76.19	74.65	75.41	91.90	85.96	83.38	68.67	87.04	84.97	85.99
FMIT	86.77	83.93	64.88	42.97	75.11	77.43	76.25	93.14	86.52	83.93	70.90	87.51	86.08	86.79
HVPNet	-	-	-	-	73.87	76.82	75.32	-	-	-	-	85.84	87.93	86.87
MNER-QG	85.31	81.65	63.41	41.32	77.43	72.15	74.70	92.92	86.19	84.52	71.67	88.26	85.65	86.94
DebiasCL	85.97	81.84	64.02	43.38	74.45	76.13	75.28	93.46	84.15	84.42	67.88	87.59	86.11	86.84
VisualPT-MoE	-	-	-	-	76.11	75.16	75.63	-	-	-	-	86.89	87.96	87.42
DPE-MNER(ours)	**87.31**	**84.36**	**65.93**	**48.48**	**76.86**	**78.27**	**77.56**	**92.37**	**87.47**	**86.33**	**73.83**	**88.46**	**87.34**	**87.90**

Table 1: Performance comparison of different competitive text-based and multi-modal methods on two Twitter datasets.

	Setting	Twitter15	Twitter17
	DEFAULT	**77.56**	**87.90**
Ablation	w/o PMI	77.17	87.54
	w/o DMF	76.92	87.19
	w/o CRM	76.85	87.38
	w/o token-level	77.23	87.54
	w/o span-level	77.05	87.49
	w/o sentence-level	77.26	87.75

Table 2: Ablation Study. Where PMI, DMF, and CRM denotes the "Prioritized Multi-modal information Integration", "Decomposed Multi-modal Fusion" and "Cross-modal Relevance Measurement"

more powerful language model leads to better performance, which is due to the fact that multimodal Named Entity Recognition (NER) is primarily text-driven. Second, we find that models incorporating multi-modal features often outperform their corresponding text-only baselines, thus demonstrating their effectiveness for this task. Third, when comparing models with similar model structures, those that fuse finer-grained representations (HVPNeT) tend to have superior results and models with smaller modality gaps in their multi-modal representations (ITA) also perform better. Finally, we evaluate the performance of BERT under three different paradigms. We observe that the span-based and CRF-based approaches yield similar results, while the iterative reasoning framework outperforms them. Significantly, our method outperforms existing text and multimodal baselines, highlighting our approach's advantages in dynamically orchestrating diverse features during iterative reasoning steps. We also compare with large language models such as ChatGPT and GPT-4. It is observed that these two LLMs underperform even the LSTM-based methods; We believe this is because MNER is a challenging task for the zero-shot reasoning of LLM.

3.4. Ablation Study

Here, we remove the essential components to observe performance changes. The results are displayed in Table 2.

Effects of Prioritized multi-modal information integration Here, we investigate the importance of the "Prioritize" strategy. We remove the "coarse-to-fine" and "easy-to-head" prioritizations in the modality fusion stage and replace them with a simple concatenation strategy. As we can see in the results, this leads to performance degradation, indicating that multi-modal fusion with prioritization makes multi-modal fusion more effective.

Effects of Decomposed multi-modal fusion To investigate the importance of the "Decomposition" strategy, we replace the hierarchical modality fusion structure with a flat fusion structure. The priority remains the same. As we can see, although this still outperforms the baseline BERT-Iterative, it suffers a more significant performance drop than removing the Prioritized multi-modal information integration. This indicates the importance of decomposing multi-modal fusion when facing diverse multi-modal representations. In further analysis, we provide a more detailed analysis of the multi-modal fusion strategies.

Effects of Cross-modal relevance measurement Finally, in analyzing the "Eliminate" strategy, we remove the losses for our image-text relevance scores, making the model learn the previously computed image-text relevance implicitly. We observed a significant drop in performance, which can be attributed to the difficulty of implicitly learning image-text relevance. We also analyzed three losses separately, and the results indicated that the most important one is span-level relevance. This is probably because it can capture entity-level relevance between text and images and thus better filter out irrelevant information. All of these three losses are important because they are used to measure the image-text relevance from different granularities, which is more effective in irrelevance filtering.

	Setting	Twitter15	Twitter17
	DEFAULT	**77.56**	**87.90**
Strategy	Max-pool	76.79	87.02
	Mean-pool	76.50	86.70
	MLP	76.28	86.34
	MoE	77.15	87.26

Table 3: Comparions with different static multi-modal fusion strategies

3.5. Further Analysis

Comparions with different static multi-modal fusion strategies In Table 3, we compare the default strategy with four static multi-modal fusion strategies, where the multi-modal fusion is pre-computed and holds for all the steps (the designs of the multi-modal fusion over spans are removed): 1) *Mean-pool*, we mean pool the image representations and fuse them with text using attention mechanisms, 2) *Max-pool*, we max pool the image representations and fuse them with text using attention mechanisms 3) *MLP*, we concatenate all the image representations and fuse them with text using an MLP layer consisting of two linear layers. 4) *MoE*, we use MoE according to VisualPT-MoE (Xu et al.). From the results, we can observe that: 1) whether *Max-pool*, *Mean-pool*, or *MLP* underperforms *MoE* and the default strategy, which is consistent with the conclusion of previous research that different modalities should be fused differently. 2) Compared to all these static fusion methods, our proposed dynamic fusion framework outperforms all of them, demonstrating that dynamic integration of multi-modal information in multiple steps can better integrate diverse multi-modal representations.

Analysis on the incorporation of image information In Table 4, we analyze the incorporation of image features from two aspects: 1) The importance of the features at different granularities, we ablate either coarse- or fine-grained image features, and to maintain the model structure, we replace the original image features with randomly initialized vectors. From the results, we can observe that removing each feature leads to performance degradation, with fine-grained features being more critical than coarse-grained features. 2) The importance of feature fusion prioritizations for different image features. We conduct three additional experiments in which the prioritizations are reversed. The results demonstrate the importance of prioritizations and the effectiveness of the prioritizations we adopted.

Scheduler	Setting	Twitter15	Twitter17
	DEFAULT	**77.56**	**87.90**
Features	w/o Coarse-grained	76.97	87.29
	w/o Fine-grained	76.54	86.86
Priority	Reverse Grain	77.26	87.37
	Reverse Gap	77.42	87.51
	Reverse Grain&Gap	77.12	87.23

Table 4: Analysis of the importance of different image features and the priority of the multi-modal information integration.

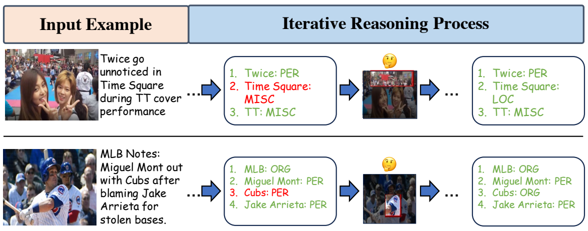

Figure 3: Two examples demonstrating the iterative reasoning process of DPE-MNER. We outline the image details that help correct the predictions in red boxes.

3.6. Case study

In Fig 3, we use two representative cases to illustrate how our framework iteratively refines the predictions with the incorporation of multi-modal information. As we can see, both cases get correct predictions on the easy entities, while the "Time Square" in the first case and "Cubs" in the second case are ambiguous and mispredicted. DPE-MNER incorporates the multi-modal information and gets the correct prediction.

4. Related Works

Multi-modal Named Entity Recognition (MNER) was first proposed by Zhang et al. (2018), aiming to improve the performance of NER using image information. There are two main lines of MNER research: exploring more effective multi-modal representations and fusing the multi-modal representations. For the first line, researchers have explored coarse-grained image representations (Yu et al., 2020) and fine-grained image representations (Zhang et al., 2021a; Chen et al., 2022c).

Some researchers attribute the bottleneck in modality fusion to the modality gap. Therefore, they directly translate the image into text to eliminate the modality gap (Wang et al., 2021). For the second line, the methods for multi-modal fusion evolve dynamically in accordance with the advancements in the mainstream field of artificial intelligence. Earlier MNER methods usually use one- or two-layer attention mechanisms to fuse image and text representations (Yu et al., 2020; Zhang et al., 2018). Then, some works try to model multi-modal fusion in a fine-grained way, such as UMGF, which uses GNN for multi-modal fusion (Zhang et al., 2021a). Some recent works try to project image representations into text spaces with a transformation matrix, then use them as prefixes for pre-trained language models, trying to fuse multi-modal information with the multi-layered transformers (Chen et al., 2022c).

There are also some branch lines on MNER, such as improving the fine-grained cross-modal alignment (Jia et al., 2023), filtering the noise in images (Yu et al., 2020), and incorporating external knowledge (Li et al., 2023).

Our work focuses on fusing all these different multi-modal representations and incorporating them reasonably to improve MNER. The most relevant to our work is VisualPT-MoE (Xu et al.). Unlike them, we leverage not only various text representations but also various image representations, and our framework is iterative. As demonstrated by extensive experiments, DPE-MNER is a stronger solution to this problem.

5. Conclusion

In this paper, we aim to fully utilize various multimodal representations in MNER for better recognition performance. To achieve this, we propose an iterative reasoning framework DPE-MNER. DPE-MNER simplifies the incorporation of these diverse representations by breaking down MNER into multiple steps. During this process, multi-modal representations are dynamically fused and integrated with a “decompose, prioritize, and eliminate” strategy. We perform extensive experiments and demonstrate the effectiveness of DPE-MNER.

Acknowledgements

The research in this article is supported by the National Key Research and Development Project (2021YFF0901602), the National Science Foundation of China (U22B2059, 62276083), and Shenzhen Foundational Research Funding (JCYJ20200109113441941), Major Key Project of PCL (PCL2021A06). Ming Liu is the corresponding author.

6. Bibliographical References

F. Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *ArXiv*, abs/2306.14122.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022a. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 118–126.

Shoufa Chen, Pei Sun, Yibing Song, and Ping Luo. 2022b. Diffusiondet: Diffusion model for object detection. *ArXiv*, abs/2211.09788.

Sijia Chen and Baochun Li. 2023. Language-guided diffusion model for visual grounding. *ArXiv*, abs/2308.09599.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022c. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv preprint arXiv:2205.03521*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

Jonathan Ho, Ajay Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239.

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. MNER-QG: An End-to-End MRC Framework for Multimodal Named Entity Recognition with Query Grounding. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 37:8032–8040.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*.

Jinyuan Li, Han Li, Zhufeng Pan, and Gang Pan. 2023. Prompt chatgpt in mner: Improved multimodal named entity recognition method based on auxiliary refining knowledge from chatgpt. *ArXiv*, abs/2305.12212.

Siyi Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *ACL*, pages 1990–1999.

Junyu Lu, Dixiang Zhang, and Pingjian Zhang. 2022. Flat multi-modal interaction transformer for named entity recognition. *arXiv preprint arXiv:2208.11039*.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, E. Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. Based Syst.*, 161:124–133.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *ArXiv*, abs/2107.00135.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Yongliang Shen, Kaitao Song, Xuejiao Tan, Dong Sheng Li, Weiming Lu, and Yue Ting Zhuang. 2023. Diffusionner: Boundary diffusion for named entity recognition. In *Annual Meeting of the Association for Computational Linguistics*.

Robert J. Sternberg and Peter A. Frensch. 1992. Complex problem solving : Principles and mechanisms. *American Journal of Psychology*, 105:501.

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*, volume 35, pages 13860–13868.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Ita: Image-text alignments for multi-modal named entity recognition. *ArXiv*, abs/2112.06482.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022. Cat-mner: Multimodal named entity recognition with knowledge-refined cross-modal attention. In *ICME*, pages 1–6.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Yanghua Xiao, and Xin Lin B. Framework with Mixture-of-Experts for Multimodal Information Extraction. 4:544–554.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: A general matching and alignment framework for multimodal named entity recognition. In *WSDM*, pages 1215–1223.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*.

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*, pages 3342–3352.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *AAAI*, volume 35, pages 14347–14355.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *AAAI*, volume 32.

Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. Reducing the Bias of Visual Objects in Multimodal Named Entity Recognition. *WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 958–966.

Gang Zhao, Guanting Dong, Yidong Shi, Haolong Yan, Weiran Xu, and Si Li. 2022. Entity-level Interaction via Heterogeneous Graph for Multimodal Named Entity Recognition. pages 6374–6379.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159.