

# Enhancing Grounded Multimodal Named Entity Recognition with Dual-Level Representation Alignment.

Anonymous

Anonymous Institution

**Abstract.** Grounded Multimodal Named Entity Recognition (GMNER) aims to identify named entities, entity types, and link them to corresponding visual objects. Existing methods face challenges in aligning entities with visual objects because of the inherent semantic differences between textual and visual features. To address this challenge, we introduce a vision-language pre-training framework named Dual-Level Representation Alignment Module (DLRAM) which is designed for entity-related tasks. Specifically, our approach performs representation alignment on two levels: the text-image level and the entity-object level. At the entity-object level, contrastive learning is employed to align specific entities with their corresponding visual objects. At the text-image level, we design a reflection mechanism to emphasize entity-related information in the image. The reflection mechanism applies object-oriented masks to guide the visual encoder in focusing on significant object features even if the target region occupies only a small area. Furthermore, given the existence of irrelevant text-image pairs, we employ the text-to-image (T2I) models to generate text-relevant images. The generated images assist in assessing the relevance of the original text-image pair and augmenting visual information. Experimental results on the GMNER-Twitter dataset demonstrate that our method outperforms existing state-of-the-art models.

**Keywords:** Grounded Multimodal Named Entity Recognition · Contrastive Learning · Visual Masking and Reflection Mechanism · Text-to-Image Models · Vision-Language Pre-Training.

## 1 Introduction

Grounded Multimodal Named Entity Recognition (GMNER) is an emerging task in information extraction. Its goal is to identify entities and their types in text, and to associate these entities with corresponding visual objects in image. GMNER produces structured triplets (entity, type, object), which are used to construct multimodal knowledge graphs (MMKGs)[4, 17, 24]. MMKGs integrate textual information with grounded visual evidence. This integration provides a reliable source of structured knowledge that enhances recommendation systems[15] and large language models[6].

Existing research on GMNER is broadly categorized into two paradigms. The first paradigm, represented by models such as RiVEG[10], divides the GMNER task into a multi-stage pipeline: (1) entity recognition, (2) visual entailment, and (3) visual grounding. However, a major drawback of this pipeline approach lies in its lengthy process, leading to inefficiency and increasing the risk of error propagation. The second paradigm employs end-to-end paradigm, such as H-Index[21], to simultaneously generate the complete triplet. In end-to-end GMNER approaches, the key challenge lies in aligning textual entities with their corresponding visual object. This task requires advanced understanding and precise alignment of cross-modal features. Yu et al.[21] addressed this challenge by utilizing implicit cross-attention mechanism. Although this approach ensures architectural simplicity, it hinders the accurate alignment of visual objects and entities.

To facilitate explicit cross-modal alignment and boost the performance of GMNER which is an entity-related task, we introduce a novel pre-training framework Dual-Level Representation Alignment Module (DLRAM). Our method performs cross-modal feature alignment along two distinct levels: the text-image level and the entity-object level. Inspired by CLIP [14], we utilize contrastive learning to train both the textual and visual encoders, enabling data from the two modalities to be projected into a unified feature space. At the text-image level, unlike conventional methods that directly align texts and images, we propose a representation alignment strategy tailored for entity-related tasks. Images often contain target regions of varying sizes. In our approach, we generate modified text-image pairs by masking these target regions. The visual encoder is employed to encode both the original and masked images. Through contrastive learning, we optimize the text and visual encoders to highlight the existence of visual targets within the global features, thereby preventing the encoder from neglecting small visual targets. For GMNER tasks, which require accurate alignment between textual entities and their corresponding visual objects, it is important to highlight target objects within the image. To meet this requirement, our method combines object detection (OD) for candidate object identification and employs contrastive learning to achieve alignment between objects and entities. This enables fine-grained cross-modal feature alignment and improves the precision of entity grounding.

As shown in Figure 1(b), our methodology aligns features across modalities at two levels. First, at the text-image level, we align the global representations produced by the textual and visual encoders. Because text-image pairs from the Internet may be irrelevant, we use text-to-image (T2I) models to synthesize images from text, thereby improving alignment performance. We evaluate the similarity between the original and synthesized images to filter out irrelevant pairs and supplement relevant visual information. Second, at the entity-object level, we perform fine-grained alignment between textual entities and their associated visual objects. For entities without corresponding visual targets, the T2I model is leveraged to enrich the visual context.

The main contributions of this paper are threefold:

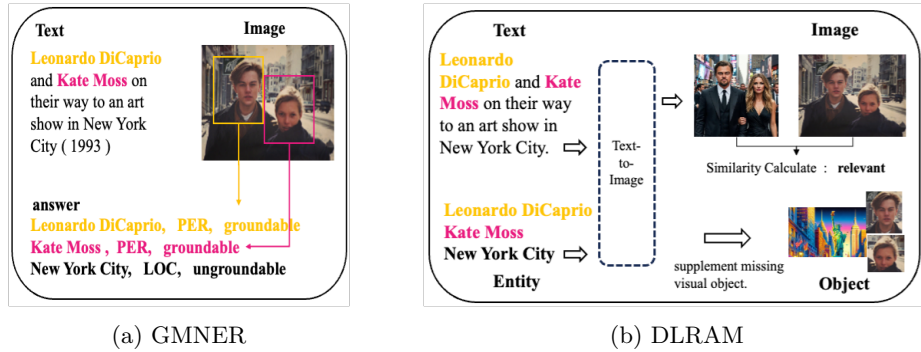


Fig. 1: Overview of GMNER and our Dual-Level Representation Alignment Module (DLRAM). Our approach aligns textual and visual modalities across two levels while utilizing text-to-image models to supplement missing visual information.

- We propose a dual-level representation alignment module for entity-related GMNER tasks, which employs two hierarchical levels of text-image level and entity-object level. At text-image level, a masking-based visual reflection mechanism is introduced to highlight corresponding targets within global visual features.
- We employ T2I model to evaluate the relevance of text-image pairs. This approach allows us to effectively remove unrelated data and supplement text-relevant images and visual objects for entities that cannot be grounded.
- Experimental results on the GMNER-Twitter dataset indicate that our method achieves improvements in performance. Moreover, ablation studies validate the impact of cross-modal feature alignment across both levels and the effectiveness of incorporating the T2I model.

## 2 Related Work

### 2.1 Multimodal Named Entity Recognition

As the application of knowledge graphs(KGs) continue to expand, research on knowledge graph construction has been steadily increasing. Multimodal Named Entity Recognition (MNER) is one of the key tasks which leverages visual information to assist in extracting and classifying textual entities. Existing MNER methods mainly focus on exploit the effective visual information and discard invalid information.

Early studies [13, 20] directly used the entire image or divided the image into multiple regions evenly. These approach to processing visual information is crude. UMGF[23] connects relevant entities across textual and visual modalities to construct a graph and perform graph-based multimodal fusion to obtain an accurate representation of text-image pairs. HvPNet[3] extracts pyramid-shaped

visual features and employs a dynamic gating module to calculate which scale of visual information should be applied to which Transformer layer.

Moreover, some methods opt to transform data from different modalities into a unified modality to bridge the gap between modalities. TMR[26] utilize diffusion-based models for text-to-image synthesis is a promising approach. PGM[9] transform images into corresponding captions and concatenating captions with the original text for entity recognition.

## 2.2 Grounded Multimodal Named Entity Recognition

Existing MNER approaches primarily focus on the extraction of textual entities. However, only extracting textual entities is helpless for constructing MMKGs. To address these limitations, Yu et al.[21] propose the GMNER task, which aims to extract the entities, their types and their corresponding visual object in image. Yu et al. collect multimodal posts on social media and build the Twitter-GMNER dataset to evaluate the model’s performance on GMNER.

There are two main paradigms for GMNER: the pipeline approach and the end-to-end approach. Under the pipeline paradigm, Yu et al. decompose GMNER into two sub-tasks: MNER, which extracts entities from text, and visual grounding (VG), which localizes entities in images. Based on this framework, RiVEG[10] adds another sub-task, Visual Entailment(VE), between MNER and VG to judge whether an entity is localizable in the image. In this approach, error propagation across sub-tasks and the increased time cost introduced by the three-stage process will impact the performance of GMNER.

For the end-to-end approach, Yu et al. propose H-Index framework to extract entity-type-object triples. Inspired by [19], they formulate the GMNER task as an index generation task. H-Index uses a seq2seq model BART[8] to encode the textual and visual inputs and decode the indexes of entities, types, and groundable or ungroundable triples. Wang et al.[1] aligns visual and textual features through pre-training, thereby demonstrating the effectiveness of aligning cross-modal features in the GMNER. Nevertheless, the alignment procedure overlooks whether the target regions are represented within the global image features, thereby restricting its effectiveness in improving entity grounding.

## 3 Methodology

As depicted in Figure 2, our model incorporates a Dual-Level Representation Alignment Module (DLRAM) to enhance semantic consistency between textual and visual representations. This module operates across two levels: text-image level and entity-object level. For the text-image alignment, we align the global features of the input text-image pairs which utilize the visual reflection mechanism to highlight the features of objects. For the entity-object alignment, we performs finer-grained alignment between entities and their corresponding objects. Furthermore, to handle potentially irrelevant text-image pairs, we employ a Text-to-Image (T2I) model to generate supplement image from the text prior to the representation alignment.

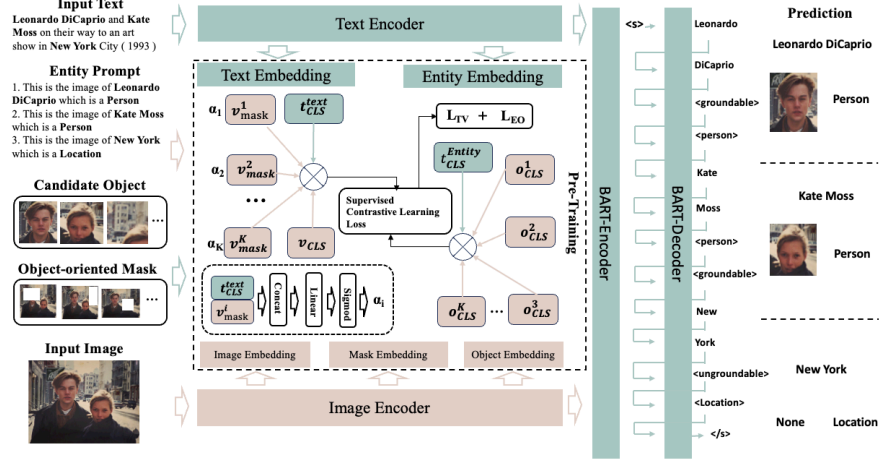


Fig. 2: An illustration of the DLRAM. For text-image alignment, we employ a masking-based visual reflection mechanism to emphasize the target visual regions within the global visual information. In the entity-object alignment, we explicitly perform fine-grained feature alignment.

### 3.1 Feature Extraction

**Textual Representation.** Given the input text  $\mathbf{s} = (s_1, s_2, \dots, s_N)$ , we use BERT[5] as textual encoder to obtain the text representations  $\mathbf{t} = \text{BERT}(\mathbf{s})$ . Here,  $\mathbf{t} = \{t_{CLS}, t_1, t_2, \dots, t_N\} \in \mathbb{R}^{d_t \times (N+1)}$  is the representation of the input text with hidden embedding size  $\mathbf{d}_t = 768$ . The representation of the special token  $[CLS]$  is considered as the textual representation, denoted as  $t_{CLS}$ .

**Visual Representation.** Given the input image  $\mathbf{v}$ , we use an object detection model VinVL[25] to recognize all candidate objects in the image. Then, we choose the top-K candidate on their detection probabilities. Finally, the image and candidate objects are encoded using a Vision Transformer ViT-B/16[7], yielding the regional embeddings denoted by  $\mathbf{v} = \{v_{CLS}, v_1, v_2, \dots, v_K\} \in \mathbb{R}^{d_v \times (K+1)}$ , where  $\mathbf{d}_v = 768$ . Similarly, The representation of the special token  $[CLS]$  is considered as the visual representation, denoted as  $v_{CLS}$ .

### 3.2 Dual-Level Representation Alignment Module

To align the textual representation  $\mathbf{t}$  and the visual representation  $\mathbf{v}$ , we propose a dual-level cross-modal alignment module (DLRAM) that aligns features from both modalities at two distinct levels: the text-image level and the entity-object level. Inspired by CLIP [14], we employ contrastive learning to align the representations of textual and visual modalities. In this framework, for each input text-image pair, we generate one positive example and  $N - 1$  negative examples at both levels.

**Text-Image Alignment** The text-image alignment focuses on aligning the overall semantic representations of the text and the image, ensuring consistency in the global context across both modalities.

Given that GMNER is an entity-related task, it is essential that the visual features extracted from the image adequately represent the objects corresponding to entities. To overcome the problem of small objects being ignored by the visual encoder, we introduce a masking-based visual reflection mechanism. For each entity in the annotated dataset, we generate a target mask to conceal its corresponding visual object. After masking, the visual encoder extracts a perturbed visual feature  $v_{mask}$ . By comparing the visual features obtained from the encoder in the presence  $v_{CLS}^i$  and absence of entity  $v_{mask}^i$ , the reflection mechanism encourages the encode to develop a more robust understanding of both the existence of entities and their features within the image. We employ the similarity function  $s = g_v(v_{CLS})^T g_t(t_{CLS})$  to measure the similarity between the image and text representations. Here,  $g_v$  and  $g_t$  are linear transformations that map the *CLS* embeddings to normalized lower-dimensional representations. This, in turn, improves cross-modal representation alignment in entity-related tasks.

$$\alpha_i = \sigma(W[v_i; t_i] + b) \quad (1)$$

$$\mathcal{L}_i^{Reflect} = -\log \frac{\exp(s(t_{CLS}^i, v_{CLS}^i)/\tau)}{\sum_{j=1}^N \exp(s(t_{CLS}^i, v_{mask}^j)/\tau)} \quad (2)$$

$$L_i^{TI} = \alpha_i \cdot \mathcal{L}_i^{Reflect} \quad (3)$$

Here,  $W$  denotes learnable weight vector,  $b$  is the learnable bias, and  $\sigma(\cdot)$  is the Sigmoid activation that outputs the relevance score between original image and masked image.

**Entity-Object Alignment** To achieve more precise alignment between textual entities and visual entities, we propose a fine-grained entity-object alignment module that aligns the representations from both modalities.

According to the annotations in the dataset, the visual objects corresponding to the entities are selected as positive samples. We follow the practice in visual grounding [22] that select  $N - 1$  box groundings from the visual regions identified by VinVL which the Intersection over Union(IoU) score between positive example are less than 0.5 as negative examples. To maximize the alignment between visual and textual modal, we introduce a contrastive loss function.

The text-image contrastive loss for the  $i^{th}$  pair is defined as follows:

$$\mathcal{L}_i^{EO} = -\log \frac{\exp(s(e_{CLS}^i, o_{CLS}^i)/\tau)}{\sum_{j=1}^N \exp(s(e_{CLS}^i, o_{CLS}^j)/\tau)} \quad (4)$$

where  $\tau$  is the temperature parameter.

Overall, the loss  $L$  during the pre-training stage can be calculated as follows:

$$\mathcal{L}_{pre-training} = \mathcal{L}^{TI} + \mathcal{L}^{EO} \quad (5)$$

### 3.3 Auxiliary Visual Information Generation Module

During the alignment process, images in the positive examples should be relevant to the text that ensure the effectiveness of cross-modal feature alignment. However, a lot of text-image pairs in posts are irrelevant to each other in social media. In Twitter-GMNER dataset, 60% entities do not have a visual object in corresponding images. Therefore, before performing DLRAM, it is necessary to determine the relevance between text-image pair.

Inspired by TMR[26], We adopt the text as the source language and leverage a Text-to-Image (T2I) model to generate images, which serve as the target language. In our work, we use DALL E[2] as T2I model. DALL E is a powerful T2I model that can produce high-quality images based on text descriptions.

The prompts for DALL E based on the text and entities are as follows:

|  |
|--|
| <p><b>Entity:</b> { Entity }</p> <p><b>Question:</b> A realistic image of a/an [Entity], clearly depicted on a plain background, showing typical features of the [Entity] with natural lighting.</p>                             |
| <p><b>Text:</b> { Text }</p> <p><b>Question:</b> An detailed illustration of the scene described by the sentence: [Text]. The image should represent the sentence in a realistic style, focus on the main action or objects.</p> |

Subsequently, we compare the features of the original image  $v$  with those of the generated image  $v_{T2I}$ . If the cosine similarity is greater than a predefined threshold  $\theta$ , we consider the text-image pair to be relevant. Otherwise, the original text-image pair is treated as irrelevant, and we substitute the original image with the generated one to perform text-image alignment. Besides, for ungroundable entities, we leverage Text-to-Image models to generate corresponding visual objects, thereby supplementing the visual information.

### 3.4 Fine-Tuning on GMNER

To extract the (entity, type, object) triples, we used the seq2seq model BART[8] for the multimodal encoder and decoder, following H-Index[21].

After encoding the input text-image pairs, we leverage the aligned representations for sequence generation. Specifically, we concatenate the textual representations  $T$  and the visual representations  $V$  and representations of visual objects  $O$  as the input for the BART encoder as follows:

$$H^e = \text{BART}_{Encoder}([T; V; O]), \quad (6)$$

Then, we utilize  $H^e$  as the input of BART decoder. At the  $i_{th}$  time step, the decoder takes  $H^e$  and the previous decoder output  $h_{i-1}$  as inputs to predict the output probability distribution  $p(y_i)$ .

$$h_i = \text{BART}_{Decoder}(H^e; h_{i-1}) \quad (7)$$

$$\bar{\mathbf{H}}_T^e = (\mathbf{T} + \text{MLP}(\mathbf{H}_T^e))/2 \quad (8)$$

$$p(\mathbf{y}_i) = \text{Softmax}([\mathbf{C}; \bar{\mathbf{H}}_T^e] \cdot h_i) \quad (9)$$

Here, *MLP* refers to a multi-layer perceptron,  $\mathbf{C} = \text{TokenEmbed}(c)$  refers to the embeddings of two indicator indexes, four entity type indexes, and special tokens  $\langle /s \rangle$ .  $\bar{\mathbf{H}}_T^e$  is a vector of length  $\mathbf{n}$ , representing each word in the sentence.

Overall, the BART decoder outputs a sequence of triplets: (entity, type, groundable/ungroundable). We employ the cross-entropy loss to optimize our seq2seq model:

$$\mathcal{L}^T = -\frac{1}{BM} \sum_{j=1}^N \sum_{i=1}^M \log p(\mathbf{y}_i^j) \quad (10)$$

where B and M denote batch size and the length of output index sequence.

Additionally, for entities whose predicted indicator is  $\langle \text{groundable} \rangle$ , we feed the output token  $h_k$  into a softmax function to obtain the probability distribution over all the candidate visual objects as following equations:

$$\bar{\mathbf{H}}_V^e = (\mathbf{V} + \text{MLP}(\mathbf{H}_V^e))/2 \quad (11)$$

$$p(\mathbf{z}_i) = \text{Softmax}(\bar{\mathbf{H}}_V^e \cdot h_k) \quad (12)$$

$\bar{\mathbf{H}}_V^e$  is a vector of length K, representing the candidate objects.

To improving the ability of visual grounding, we minimize the KL Divergence loss between the predicted region distribution  $p(z_i)$  and the region supervision  $g(z_k)$ :

$$\mathcal{L}^V = -\frac{1}{BE} \sum_{j=1}^N \sum_{k=1}^E g(\mathbf{z}_k^j) \log \frac{p(\mathbf{z}_i^j)}{g(\mathbf{z}_k^j)} \quad (13)$$

where E is the number of groundable entities.

Finally, we combine  $\mathcal{L}^T$  and  $\mathcal{L}^V$  as our loss of fine-tuning:

$$\mathcal{L}_{fine-tuning} = \mathcal{L}^T + \mathcal{L}^V \quad (14)$$

## 4 Experiment

### 4.1 Experimental Setup

In our method, we utilize VinVL to detect the top-k visual regions with the highest confidence scores as candidate visual objects, with k set to 16. In contrastive learning, the parameter  $N$  is set to 16. During training, we use the AdamW optimizer for parameter tuning. The batch size and training epoch are set to 32 and 20. For the learning rate, we set  $3e-5$ . In Section 3.3, when assessing the relevance between images and text, we set the  $\theta$  to 0.6.

## 4.2 Datasets

To assess the effectiveness of DLRAM framework, we performed experiments on the popular Twitter-GMNER [21] dataset. The Twitter-GMNER dataset consists of 7,000/1,500/1,500 text-image pairs in the train/dev/test sets, respectively. Additionally, it contains 16,778 entities, approximately 59.6% of them do not have corresponding visual objects in the images. The entity types in Twitter-GMNER are divided into four categories: PER, LOC, ORG, and MISC.

## 4.3 Evaluation Metrics

The goal of GMNER is to extract (entity, type, object) triples from text-image pairs. GMNER can be further divided into MNER (entity and type extraction) and EEG (entity extraction and grounding). For MNER, a prediction is considered correct only if both the entity and type match the gold labels. For EEG, if the entity is groundable, a prediction is correct when the Intersection over Union (IoU) score between the predicted region and ground-truth (GT) bounding boxes is greater than 0.5. If the entity is ungroundable, predicting *None* is considered correct.

## 4.4 Experimental Benchmarks

We selected three paradigmatic approaches: NER+None, MNER+OD+VG, and GMNER.

**NER+None.** To compare the models ability to extract entities and types from text, we selected several representative NER methods: HBiLSTM-CRF-None, BERT-None, BERT-CRF-None, and BARTNER-None. These text-only baselines default the prediction of visual object is None.

**MNER+OD+VG.** This framework consists of three stages: MNER, OD, and VG. The MNER method extract textual entities and their types, while the object detection (OD) method identifies candidate object regions within the image. The Visual Grounding (VG) method is employed to link entities to their corresponding objects detected by OD. Following previous studies [16], we choose several approaches by combining MNER, OD and VG methods, including GVATT[12]-RCNN-EVG[21], UMT[20]-RCNN-EVG, UMGF[23]-VinVL-EVG, ITA[18]-VinVL-EVG, and BARTMNER-VinVL-EVG.

**GMNER.** In recent years, many studies have focused on end-to-end frameworks for GMNER. We select representative state-of-the-art methods for comparison. GMNER [21] convert data into labels and directly generate entity-type-object triples. GMDA [11] introduces a generative multimodal augmentation technique to enrich training data and enhance model performance.

Table 1: Performance comparison of different methods on GMNER, MNER and EEG tasks.

| Modal. Methods |                      | GMNER        |              |              | MNER         |       |              | EEG          |              |              |
|----------------|----------------------|--------------|--------------|--------------|--------------|-------|--------------|--------------|--------------|--------------|
|                |                      | Pre.         | Rec.         | F1.          | Pre.         | Rec.  | F1.          | Pre.         | Rec.         | F1.          |
| T-only         | HBiLSTM-CRF-None     | 43.56        | 40.69        | 42.07        | 78.80        | 72.61 | 75.58        | 49.17        | 45.92        | 47.49        |
|                | BERT-None            | 42.18        | 43.76        | 42.96        | 77.26        | 77.41 | 77.30        | 46.76        | 48.52        | 47.63        |
|                | BERT-CRF-None        | 42.73        | 44.88        | 43.78        | 77.23        | 78.64 | 77.93        | 46.92        | 49.28        | 48.07        |
|                | BARTNER-None         | 44.61        | 45.04        | 44.82        | 79.67        | 79.98 | 79.83        | 48.77        | 49.23        | 48.99        |
| T+V            | GVATT-RCNN-CRF       | 49.36        | 47.80        | 48.57        | 78.21        | 74.39 | 76.26        | 54.19        | 52.48        | 53.32        |
|                | UMT-RCNN-EVG         | 49.16        | 51.48        | 50.29        | 77.89        | 79.28 | 78.58        | 53.55        | 56.08        | 54.78        |
|                | UMT-VinVL-EVG        | 50.15        | 52.52        | 51.31        | 77.89        | 79.28 | 78.58        | 54.35        | 56.91        | 55.60        |
|                | UMGF-VinVL-EVG       | 51.62        | 51.72        | 51.67        | 79.02        | 78.64 | 78.83        | 55.68        | 55.80        | 55.74        |
|                | ITA-VinVL-EVG        | 52.37        | 50.77        | 51.57        | 80.40        | 78.37 | 79.37        | 56.57        | 54.84        | 55.69        |
|                | BARTMNER-VinVL-EVG   | 52.47        | 52.43        | 52.45        | <b>80.65</b> | 80.14 | <b>80.39</b> | 55.68        | 55.63        | 55.66        |
|                | H-Index              | 56.16        | 56.67        | 56.41        | 79.37        | 80.10 | 79.73        | 60.90        | 61.46        | 61.18        |
|                | H-Index + GMDA       | 56.27        | 57.44        | 56.85        | —            | —     | —            | —            | —            | —            |
|                | Ours H-Index + DLRAM | <b>59.43</b> | <b>60.83</b> | <b>60.12</b> | 80.00        | 78.76 | 79.37        | <b>64.51</b> | <b>64.93</b> | <b>64.71</b> |

#### 4.5 Experimental Analysis

In Table 1, we show the result of different methods on GMNER, MNER, and EEG.

**Results on GMNER.** We compared our framework with text-only methods, pipeline methods of GMNER, and end-to-end methods of GMNER. First, for text-only methods, BARTNER-None outperforms other approaches which shows the effectiveness of index generation framework in NER. However, due to the ambiguity in single-modality text data, The performance of entity extraction is not yet optimal. Second, comparing all the pipeline methods, BARTMNER-VinVL-EVG obtains the best result which primarily due to the performance of MNER subtask. Third, compared to end-to-end GMNER methods, our method achieves better results. Our method has improved by 3.71 absolute percentage points compared to the H-Index method thanks to our DLRAM. Our method aligns textual and visual features for entity-related tasks. This enables the learning of more multimodal features while mitigating the adverse effects resulting from semantic gaps.

**Results on MNER.** To assess the performance of our method in extracting entities and their types, we also compared its scores on the MNER subtask. The results show that mere alignment of textual and visual features is inadequate for effectively enhancing the recognition of entities and their types.

**Results on EEG.** To evaluate the performance of our method in entity extraction and grounding, we assessed its scores on the EEG subtask. Previous approaches have shown that simply linking textual entities with visual entities restricts the effectiveness of GMNER. In our method, we introduce cross-modal alignment specifically for entity-related tasks, mapping textual and visual fea-

tures into a shared feature space while enhancing the presence of entities within global features. According to the experimental results, our method achieves a 3.53 increase in F1 score compared to GMNER.

#### 4.6 Ablation Study.

**The effectiveness of text-image and entity-object alignment.** To verify the effectiveness of the entity-object and text-image alignments in our method, we conduct ablation experiments. One performs only entity-object alignment, while the other performs only text-image alignment. For the GMNER, when we remove entity-object alignment, we observe that the performance drop 2.66 percentage points on F1 score. This indicates that entity-object alignment enables fine-grained cross-modal alignment which allows the model to more easily understand the features of textual and visual entities, facilitating the matching of entities with their corresponding visual objects.

Table 2: Effect of Text-Image and Entity-Object alignment

| Methods          | GMNER |       |               |
|------------------|-------|-------|---------------|
|                  | Pre.  | Rec.  | F1.           |
| DLRAM            | 59.43 | 60.83 | 60.12         |
| - w/o <i>TIA</i> | 58.32 | 59.65 | 58.98 (1.14↓) |
| - w/o <i>EOA</i> | 57.31 | 57.62 | 57.46 (2.66↓) |

**The Effectiveness of T2I Model.** Irrelevance between images and texts is common in text-image pair datasets. Utilizing T2I models to evaluate the relevance of text-image pairs and to supplement missing visual information can improve the effectiveness of cross-modal feature alignment. To validate the effectiveness of employing T2I models for GMNER, we conducted experiments using None, DALL E, and DALL E 3 as the T2I models, respectively. As shown in Table 3, when the T2I model is not been used, the F1 score for GMNER is the lowest. Furthermore, compared to DALL E, DALL E 3 benefits from optimized text descriptions in its training data, allowing the generated images to more accurately reflect the content of the text. Therefore, DALL E 3 facilitates better alignment between entities and visual objects.

#### 4.7 Case Study.


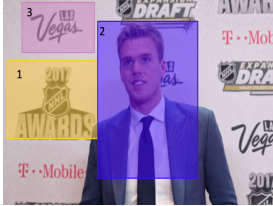
We conduct a case study to compare the predictions of *H-Index*, *DLRAM* on the test sample from the GMNER-Twitter dataset. As shown in Table 4, we found that while both of them correctly identify the entity-type pairs, their entity grounding performance varies significantly. H-Index struggles to align entities with their corresponding visual regions. In Table 4(a), H-Index failed to recognize the logos corresponding to the (*premier league*, *ORG*). Our method accurately

Table 3: Comparison of None, DALL E, DALL E 3 in complementary visual information

| Methods        | GMNER |       |               |
|----------------|-------|-------|---------------|
|                | Pre.  | Rec.  | F1.           |
| DLRAM-None     | 57.52 | 58.21 | 57.86         |
| DLRAM-DALL E   | 59.43 | 60.83 | 60.12 (2.26 ) |
| DLRAM-DALL E 3 | 59.82 | 61.23 | 60.52 (2.66 ) |

distinguishes and grounds the logos for both organizations. In Table 4(b), the entity *Hart Trophy* does not have a corresponding visual region in the image. However, *H-Index* incorrectly associate it with a visual region, whereas *DLRAM* correctly identifies the entity as ungrounded. These results indicate that DLRAM can effectively identify smaller target regions.

Table 4: A case study analyzing two cases of prediction.

| (a)  | (b)  |
|--|--|
| Leicester City (ORG, Box1) striker Jamie Vardy (PER, Box2) was named Premier League (ORG, Box3) player.        | NHL Awards (OTHER, Box1) : Connor McDavid (PER, Box2) wins Hart Trophy (ORG, N/A) as league MVP.             |
|                             |                          |
| <b>H-Index</b><br>(Leicester City, ORG, Box1)  <br>(Jamie Vardy, PER, Box2)  <br>(Premier League, ORG, Box3)   | <b>H-Index</b><br>(NHL Awards, OTHER, Box1)  <br>(Connor McDavid, PER, Box2)  <br>(Hart Trophy, ORG, Box1)   |
| <b>DLRAM</b><br>(Leicester City, ORG, Box1)  <br>(Jamie Vardy, PER, Box2)  <br>(Premier League, ORG, Box3)     | <b>DLRAM</b><br>(NHL Awards, OTHER, Box1)  <br>(Connor McDavid, PER, Box2)  <br>(Hart Trophy, ORG, N/A)      |

## 5 Conclusion

In this work, we address the challenge of the semantic gap between multimodal features, which leads to poor performance in recognizing entity-object pairs.

Our methodology introduces a dual-level alignment module (DLRAM), including text-image and entity-object levels, which employs contrastive learning to align the representations from textual and visual encoders. At the text-image level, we introduce a masking-based visual reflection mechanism that emphasizes the existence of entities in the global representation, thus preventing small target regions from being ignored. Besides, we leverage Text-to-Image models to generate relevant images and enrich the visual context. Experimental validation shows that our method improves the F1 score for the GMNER task, which highlights the importance of robust feature alignment across modalities.

## References

1. Bao, X., Tian, M., Wang, L., Zha, Z., Qin, B.: Contrastive pre-training with multi-level alignment for grounded multimodal named entity recognition. In: Proceedings of the 2024 international conference on multimedia retrieval. pp. 795–803 (2024)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. <https://cdn.openai.com/papers/dall-e-3.pdf> **2**(3), 8 (2023)
3. Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., Huang, F., Si, L., Chen, H.: Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. arXiv preprint arXiv:2205.03521 (2022)
4. Chen, Z., Zhang, Y., Fang, Y., Geng, Y., Guo, L., Chen, J., Liu, X., Pan, J.Z., Zhang, N., Chen, H., et al.: Knowledge graphs for multi-modal learning: Survey and perspective. Information Fusion p. 103124 (2025)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
6. Dong, J., Zhang, Q., Zhou, H., Zha, D., Zheng, P., Huang, X.: Modality-aware integration with large language models for knowledge-based visual question answering. arXiv preprint arXiv:2402.12728 (2024)
7. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 7871–7880 (2020)
9. Li, J., Li, H., Pan, Z., Sun, D., Wang, J., Zhang, W., Pan, G.: Prompting chatgpt in mner: Enhanced multimodal named entity recognition with auxiliary refined knowledge. arXiv preprint arXiv:2305.12212 (2023)
10. Li, J., Li, H., Sun, D., Wang, J., Zhang, W., Wang, Z., Pan, G.: Llms as bridges: Reformulating grounded multimodal named entity recognition. arXiv preprint arXiv:2402.09989 (2024)
11. Li, Z., Yu, J., Yang, J., Wang, W., Yang, L., Xia, R.: Generative multimodal data augmentation for low-resource multimodal named entity recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7336–7345 (2024)

12. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: Visual attention model for name tagging in multimodal social media. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1990–1999 (2018)
13. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity recognition for short social media posts. arXiv preprint arXiv:1802.07862 (2018)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., Wang, Z., Zheng, K.: Multi-modal knowledge graphs for recommender systems. In: Proceedings of the 29th ACM international conference on information & knowledge management. pp. 1405–1414 (2020)
16. Wang, J., Li, Z., Yu, J., Yang, L., Xia, R.: Fine-grained multimodal named entity recognition and grounding with a generative framework. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3934–3943 (2023)
17. Wang, M., Wang, H., Qi, G., Zheng, Q.: Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Research **22**, 100159 (2020)
18. Wang, X., Gui, M., Jiang, Y., Jia, Z., Bach, N., Wang, T., Huang, Z., Tu, K.: Ita: Image-text alignments for multi-modal named entity recognition. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 3176–3189 (2022)
19. Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., Qiu, X.: A unified generative framework for various ner subtasks. arXiv preprint arXiv:2106.01223 (2021)
20. Yu, J., Jiang, J., Yang, L., Xia, R.: Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics (2020)
21. Yu, J., Li, Z., Wang, J., Xia, R.: Grounded multimodal named entity recognition on social media. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9141–9154 (2023)
22. Yu, Z., Yu, J., Xiang, C., Zhao, Z., Tian, Q., Tao, D.: Rethinking diversified and discriminative proposal generation for visual grounding. arXiv preprint arXiv:1805.03508 (2018)
23. Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G.: Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14347–14355 (2021)
24. Zhang, J., Wang, J., Wang, X., Li, Z., Xiao, Y.: Aspectmmkg: A multi-modal knowledge graph with aspect-aware entities. In: Proceedings of the 32nd ACM international conference on information and knowledge management. pp. 3361–3370 (2023)
25. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2021)
26. Zheng, C., Feng, J., Cai, Y., Wei, X., Li, Q.: Rethinking multimodal entity and relation extraction from a translation point of view. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6810–6824 (2023)