LLMs as Bridges: Reformulating Grounded Multimodal Named Entity Recognition

Jinyuan Li[1], **Han Li**[2], **Di Sun**[4], **Jiahao Wang**[3], **Wenkun Zhang**[5], **Zan Wang**[1,3], **Gang Pan**[1,3,*]
[1]School of New Media and Communication, Tianjin University
[2]College of Mathematics, Taiyuan University of Technology
[3]College of Intelligence and Computing, Tianjin University
[4]Tianjin University of Science and Technology
[5]University of Copenhagen
{jinyuanli, wjhwtt, wangzan, pangang}@tju.edu.cn, lihan0928@link.tyut.edu.com
wenkun.zhang@sund.ku.dk, dsun@tust.edu.cn

Abstract

Grounded Multimodal Named Entity Recognition (GMNER) is a nascent multimodal task that aims to identify named entities, entity types and their corresponding visual regions. GMNER task exhibits two challenging properties: 1) The weak correlation between image-text pairs in social media results in a significant portion of named entities being ungroundable. 2) There exists a distinction between coarse-grained referring expressions commonly used in similar tasks (*e.g.*, phrase localization, referring expression comprehension) and fine-grained named entities. In this paper, we propose RiVEG, a unified framework that reformulates GMNER into a joint MNER-VE-VG task by leveraging large language models (LLMs) as a connecting bridge. This reformulation brings two benefits: 1) It maintains the optimal MNER performance and eliminates the need for employing object detection methods to pre-extract regional features, thereby naturally addressing two major limitations of existing GMNER methods. 2) The introduction of entity expansion expression and Visual Entailment (VE) module unifies Visual Grounding (VG) and Entity Grounding (EG). It enables RiVEG to effortlessly inherit the Visual Entailment and Visual Grounding capabilities of any current or prospective multimodal pretraining models. Extensive experiments demonstrate that RiVEG outperforms state-of-the-art methods on the existing GMNER dataset and achieves absolute leads of 10.65%, 6.21%, and 8.83% in all three subtasks. Code and datasets publicly available at https://github.com/JinYuanLi0012/RiVEG.

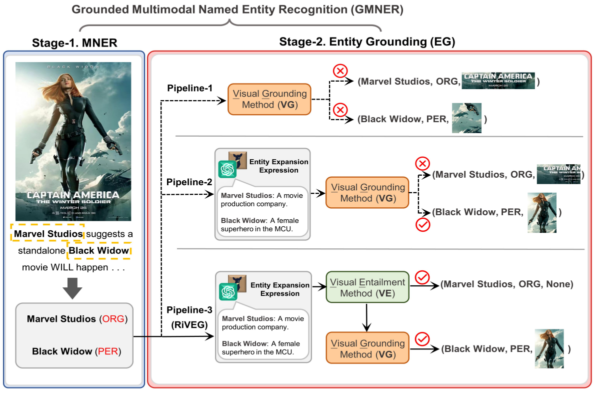

Grounded Multimodal Named Entity Recognition (GMNER)
Stage-1. MNER
Stage-2. Entity Grounding (EG)
CAPTAIN AMERICA
THE WINTER SOLDIER
Marvel Studios suggests a standalone Black Widow movie WILL happen . . .
Marvel Studios (ORG)
Black Widow (PER)
Pipeline-1
Visual Grounding Method (VG)
(Marvel Studios, ORG,)
(Black Widow, PER,)
Pipeline-2
Entity Expansion Expression
Marvel Studios: A movie production company.
Black Widow: A female superhero in the MCU.
Visual Grounding Method (VG)
(Marvel Studios, ORG,)
(Black Widow, PER,)
Pipeline-3 (RiVEG)
Entity Expansion Expression
Marvel Studios: A movie production company.
Black Widow: A female superhero in the MCU.
Visual Entailment Method (VE)
Visual Grounding Method (VG)
(Marvel Studios, ORG, None)
(Black Widow, PER,)

Figure 1: Pipeline-1 illustrates the challenges posed by GMNER task to existing Visual Grounding methods. RiVEG builds a bridge between noun phrases and named entities and introduces Visual Entailment to address the weak correlation between image and text.

1 Introduction

Multimodal Named Entity Recognition (MNER) (Lu et al., 2018a) aims to extract named entities and corresponding categories from image-text pairs sourced from social media. As one of the most important subtask in multimodal information processing, MNER has received extensive attention in recent years (Wang et al., 2022c,b; Li et al., 2023; Cui et al., 2024). Due to the pressing demand for constructing structured knowledge graphs, human-machine interaction and robotics, recent research endeavors to extend the scope of MNER (Wang et al., 2023). As a further extension, Grounded Multimodal Named Entity Recognition (GMNER) (Yu et al., 2023) aims to extract text named entities, entity types and their bounding box groundings from image-text pairs.

Analyzing GMNER from the two perspectives of MNER and EG, existing GMNER methods (Yu et al., 2023) have several significant limitations. Limitation 1. Entity Grounding (EG) is an important part in GMNER, which aims to determine the groundability and visual grounding results of text named entities predicted by the MNER model. Obviously, effective EG depends on accurate MNER. But the MNER performance of the existing multimodal fusion GMNER methods (Yu et al., 2023) is suboptimal. This may be due to the introduction of visual object features destroying the original

[*] corresponding authors.

text feature representation in MNER (Wang et al., 2022c). A series of MNER studies in the Text-Text paradigm (*i.e.*, represent images using text) (Wang et al., 2022b; Li et al., 2023) show that text representation plays the most critical role in MNER. Models with the best MNER performance typically do not involve any cross-modal interactions. However, existing end-to-end GMNER methods (Yu et al., 2023) violate this in order to obtain EG capabilities with the help of cross-modal interactions. This results in the feature representation learned by the model being neither the best for MNER nor the best for EG. Therefore, splitting the prediction and grounding of named entities into separate stages is necessary. Limitation 2. Existing GMNER methods (Yu et al., 2023; Wang et al., 2023) typically employ object detection (OD) techniques to extract features from candidate regions, which are then aligned with text entities. But as shown in Table 1, our statistical results[1] indicate that the widely-adopted OD methods exhibit unsatisfactory performance in recognizing candidate visual objects in GMNER dataset. This implies that there exists a natural upper limit to the effectiveness of existing GMNER methods. In instances where the OD methods are unable to detect the visual ground-truth bounding boxes for textual named entities, even if the MNER part predicts the correct textual named entity, the EG result will definitely be wrong. Therefore, it is necessary to avoid extracting visual features from candidate image regions.

Visual Grounding (VG) (Rohrbach et al., 2016; Yu et al., 2018) aims to correlate natural language descriptions with specific regions in images. Although VG has a similar paradigm to the Entity Grounding (EG) in GMNER, existing VG methods (Zhu et al., 2022; Deng et al., 2021; Wang et al., 2022a) exhibit significant deficiencies when applied to the GMNER dataset. This is attributed to two fundamental distinctions between these two tasks: Differentiation 1. Referring to the Pipeline-1 and Pipeline-2 in Figure 1, the text input for the EG task consists of real-world named entities, which is notably different from the referring expressions or noun phrases typically used in the VG method. There is a substantial difference that is almost non-generalizable between them. The VG method can comprehend the description "A female superhero in the MCU" but struggles with understanding named entities like "Black Widow". Therefore, a bridge between noun phrases and named entities needs to be built. Differentiation 2. Due to the social media usage patterns of users, image-text pairs from social media exhibit apparent weak correlation. It implies that the textual named entity may not necessarily be linked to a specific region in the image. Yu et al. (2023) shows that approximately 60% of the named entities in the GMNER dataset are ungroundable. But the classic VG task stipulates that the input referring expression must match an object in the image, without considering ungroundable expressions that do not match any object. It means that the behavior of the existing VG methods is undefined if the named entity does not correspond to any object in the image. As shown in Pipeline-1 and Pipeline-2 of Figure 1, existing VG methods inevitably make errors when they confront the ungroundable text input like "Marvel Studios". Our experiments demonstrate that the Prec@0.5 of the state-of-the-art VG methods applied to the GMNER dataset is 21.87% (Wang et al., 2022a) and 9.23% (Zhu et al., 2022), respectively.[2] Therefore, a separate module is needed to determine the groundability of text inputs.

OD methods	Top-10	Top-15	Top-20
Anderson et al. (2018)	59.87%	69.84%	76.11%
Zhang et al. (2021b)	66.69%	74.62%	84.29%

Table 1: TopN-Prec@0.5 scores of two widely-adopted OD methods on Twitter-GMNER dataset (Yu et al., 2023). Existing GMNER methods are significantly limited by OD techniques. Because these visual candidate areas do not necessarily include the visual gold label of Entity Grounding.

Given the above differences, the potential of VG methods cannot be unleashed in EG. Considering that most VG methods do not require pre-extraction of regional features, we ask: *Is it possible to reformulate GMNER as a two-stage union of MNER-EG? The first stage focuses on optimizing the MNER effect to solve the Limitation 1 of the existing GMNER method. The second stage uses the VG method to naturally bypass regional feature extraction to solve Limitation 2. In the second stage, to adapt the VG method to the EG task, is it*

[1] Here, TopN-Prec@0.5 refers to the accuracy metric in object detection methods across all samples. It measures the proportion of named entities where at least one of the Top-N predicted bounding boxes based on detection probability has an IoU of 0.5 or greater with the ground-truth bounding box.

[2] We fine-tune VG models using the GMNER dataset. Text inputs are named entities and entity types. For ungroundable entities, the coordinates of gold labels are defined as all zeros.

conceivable to establish a bridge between named entities and noun phrases to address Differentiation 1 between the VG method and EG task? Moreover, is it feasible to resolve Differentiation 2 by devising a sensible filtering mechanism?

In this paper, we introduce GMNER into a multi-stage framework consisting of MNER, VE and VG (RiVEG). This modular design inherently mitigates two limitations of existing GMNER methods, enabling the connection of three unrelated independent tasks in series within GMNER and unlocking significant potential. Specifically, in order to ensure the optimal performance of the MNER stage, we leverage auxiliary refined knowledge distilled from large language models (LLMs) to aid in the identification of named entities. And we introduce a Visual Entailment (VE) module to handle the weak correlation between image and text. Named entities are heuristically converted by LLMs into corresponding entity expansion expressions to facilitate the VE and VG modules to determine the groundability and grounding results of named entities. Main contributions are summarized as follows:

(i) RiVEG extends the text input of Visual Entailment and Visual Grounding from limited natural language expressions that require manual definition to infinite named entities that exist naturally by leveraging LLMs as bridges.

(ii) RiVEG increases the application scope of VG models to the cases of no right objects in image and demonstrates a new paradigm for GMNER. This paradigm supports data augmentation and facilitates subsequent upgrades.

(iii) All 14 variants of RiVEG achieve new state-of-the-art performance on the GMNER dataset. And after effective data augmentation, the MNER module of RiVEG achieves new state-of-the-art performance on two classic MNER datasets.

2 Related Work

2.1 Multimodal Named Entity Recognition

MNER is a relatively independent multimodal subtask. Early MNER methods attempt to simply interact text and images (Moon et al., 2018; Lu et al., 2018b; Yu et al., 2020). To deal with the limitations of this shallow interaction, subsequent methods explore various cross-modal fusion approaches (Zhang et al., 2021a; Wang et al., 2022e; Jia et al., 2023; Chen et al., 2022a,b). There is a trend to solve MNER through the Text-Text paradigm (Wang et al., 2022c,b; Li et al., 2023), and this paradigm generally yields better performance. Despite obtaining promising results, these methods lack any visual understanding capabilities. The GMNER task poses significant challenges to the above research.

2.2 Grounded Multimodal Named Entity Recognition and Visual Grounding

GMNER is a further exploration of MNER. Existing GMNER works (Yu et al., 2023; Wang et al., 2023) aim to utilize OD methods for obtaining region proposals and subsequently sorting them based on their relevance to named entities. Such methods are inevitably limited by OD methods. Visual Grounding (VG) is a representative multimodal subtask. The progress in multimodal pretraining research (Zhu et al., 2022; Wang et al., 2022a; Yan et al., 2023) continues to contribute to the improvement of VG methods. However, the VG task has some strong predefined constraints. It presupposes that there are objects in the image that match the input text query. This renders VG methods inapplicable to the GMNER task that deviates from such predefined constraints. In this paper, we introduce the VE module to reconcile this conflict. Our method uses the VG method as the main body of visual understanding and achieves better performance than the existing GMNER methods.

3 Methodology

RiVEG is mainly divided into two stages. In the stage of Named Entity Recognition and Expansion, we employ auxiliary refined knowledge obtained from LLMs to ensure optimal MNER performance. Furthermore, in order to adapt existing VG methods to EG task, LLMs is considered as a bridge between named entities and referring expressions. RiVEG instructs LLMs to convert named entities into named entity referring expressions (detailed in §3.2). In the stage of Named Entity Grounding, RiVEG reformulates the entire EG task as a union of VE and VG. To reconcile the challenge between the weak correlation of image-text pairs in GMNER and the strong predefined constraints of existing VG methods, we design the VE module for buffering. Images and named entity referring expressions serve as the input to VE module for determining the groundability of the named entities. Finally, the named entities deemed groundable

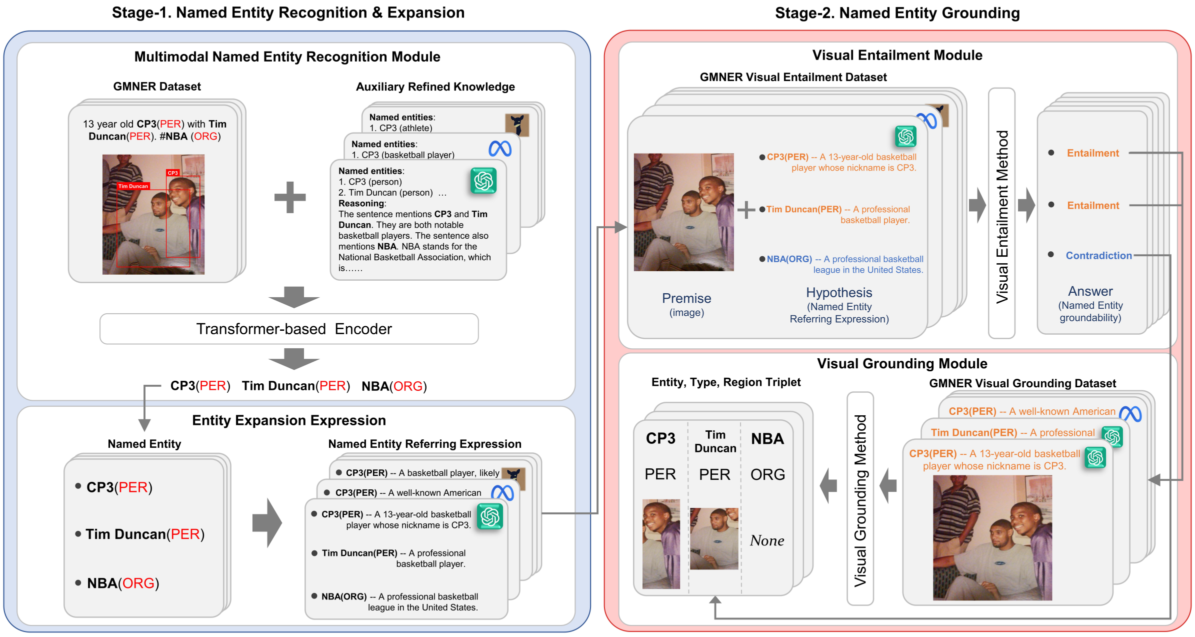

Stage-1. Named Entity Recognition & Expansion
Multimodal Named Entity Recognition Module
GMNER Dataset
13 year old CP3(PER) with Tim Duncan(PER). #NBA (ORG)
CP3
Tim Duncan
Auxiliary Refined Knowledge
Named entities:
1. CP3 (athlete)
Named entities:
1. CP3 (basketball player)
Named entities:
1. CP3 (person)
2. Tim Duncan (person) …
Reasoning:
The sentence mentions CP3 and Tim Duncan. They are both notable basketball players. The sentence also mentions NBA. NBA stands for the National Basketball Association, which is……
Transformer-based Encoder
CP3(PER) Tim Duncan(PER) NBA(ORG)
Entity Expansion Expression
Named Entity
• CP3(PER)
• Tim Duncan(PER)
• NBA(ORG)
Named Entity Referring Expression
• CP3(PER) -- A basketball player, likely
• CP3(PER) -- A well-known American
• CP3(PER) -- A 13-year-old basketball player whose nickname is CP3.
• Tim Duncan(PER) -- A professional basketball player.
• NBA(ORG) -- A professional basketball league in the United States.
Stage-2. Named Entity Grounding
Visual Entailment Module
GMNER Visual Entailment Dataset
• CP3(PER) -- A 13-year-old basketball player whose nickname is CP3.
• Tim Duncan(PER) -- A professional basketball player.
• NBA(ORG) -- A professional basketball league in the United States.
Premise (image)
Hypothesis (Named Entity Referring Expression)
Visual Entailment Method
• Entailment
• Entailment
• Contradiction
Answer (Named Entity groundability)
Visual Grounding Module
Entity, Type, Region Triplet
CP3
Tim Duncan
NBA
PER
PER
ORG
None
Visual Grounding Method
GMNER Visual Grounding Dataset
CP3(PER) -- A well-known American
Tim Duncan(PER) -- A professional
CP3(PER) -- A 13-year-old basketball player whose nickname is CP3.

Figure 2: The architecture of RiVEG.

alongside images are employed as inputs to VG module to obtain the final grounding results (detailed in §3.3). We also design a data augmentation method suitable for RiVEG. An overview of the RiVEG is depicted in Figure 2.

3.1 Task Formulation

Given a sentence $S = \{s_1, \cdots, s_n\}$ with n tokens and its corresponding image I, the objective of GMNER is to identify and categorize named entities (*i.e.*, PER, LOC, ORG and $MISC$) within the sentence while concurrently determining the visually grounded region corresponding to each entity. Formulate that the output of GMNER comprises a set of multimodal entity triples:

$$Y = \{(e_1, t_1, v_1), \cdots, (e_n, t_n, v_n)\}$$

where e_i is one of the entities in sentence, t_i refers to the type of e_i, and v_i represents the corresponding visually grounded region. Note that if e_i is ungroundable, v_i is specified as $None$; otherwise, v_i consists of a 4D coordinates representing the top-left and bottom-right locations of the grounded bounding box, *i.e.*, $(v_i^{x1}, v_i^{y1}, v_i^{x2}, v_i^{y2})$.

3.2 Stage-1. Named Entity Recognition and Expansion

Multimodal Named Entity Recognition Module Incorporating document-level extended knowledge into original multimodal samples has been proven to be one of the most effective methods for MNER (Wang et al., 2022c,b; Li et al., 2023). PGIM (Li et al., 2023) points out that despite the inherent limitations of generative models in sequence labeling tasks, LLM serve as an implicit knowledge base that is highly suitable for MNER task. We use the auxiliary refined knowledge[3] provided by PGIM to enhance the predictive performance of named entities. And we additionally use the same algorithm to generate different auxiliary refined knowledge for the same sample from more versions of LLMs to perform data augmentation. Specifically, definite auxiliary refined knowledge retrieved by LLMs as $Z = \{z_1, \cdots, z_m\}$. We take the concatenation of the original sentence $S = \{s_1, \cdots, s_n\}$ together with the auxiliary refined knowledge Z as the input of the transformer-based encoder:

$$\text{embed}([S; Z]) = \{r_1, \cdots, r_n, \cdots, r_{n+m}\}$$

And the acquired token representations $R = \{r_1, \cdots, r_n\}$ is fed into a linear-chain Conditional Random Field (CRF) (Lafferty et al., 2001) to generate the predicted sequence $y = \{y_1, \cdots, y_n\}$:

$$P(y|S, Z) = \frac{\prod_{i=1}^{n} \psi(y_{i-1}, y_i, r_i)}{\sum_{y' \in Y} \prod_{i=1}^{n} \psi(y'_{i-1}, y'_i, r_i)}$$

[3] Original text and image captions are used as inputs to LLMs, and LLMs are guided to generate extended explanations of the original samples.

where ψ is the potential function and Y is the set of all possible label sequences given the input S and Z. Finally, given the input sequence with gold labels $\boldsymbol{y}^*$, we employ the negative log-likelihood (NLL) loss function for training model:

$$\mathcal{L}_{\text{MNER}}(\theta) = -\log P_\theta(\boldsymbol{y}^*|S, Z)$$

Note that the approach employed by this module is not unique and can be replaced or upgraded by any more powerful MNER methods in the future.

Entity Expansion Expression After extracting named entities in the sentence, RiVEG further guides LLMs to reformulate the fine-grained named entities into coarse-grained entity expansion expressions. Specifically, for each extracted named entity, we formulate it as the following template:

Background: $\{Image\ Caption\}$
Text: $\{Sentence\}$
Question: `In the context of the provided information, tell me briefly what is the` $\{Named\ Entity\}$ `in the Text?`
Answer: $\{Entity\ Expansion\ Expression\}$

where $\{Entity\ Expansion\ Expression\}$ is reserved for LLMs to generate. A limited number of predefined in-context examples are employed to guide the format of responses (detailed in Appendix Figure 5). Finally, we concatenate named entities, entity categories, and entity expansion expressions into the final named entity referring expressions: $Named\ Entity(Entity\ Categoriy) - Entity\ Expansion\ Expression$

Furthermore, in order to perform data augmentation, we use different LLMs to generate different entity expansion expressions of the same entity.

3.3 Stage-2. Named Entity Grounding

Through the aforementioned transformation, the GMNER task is naturally reframed as the MNER-VE-VG task. Notably, a large number of well-established studies exist on MNER, VE, and VG. The VE and VG methods that can be utilized at this stage are also not unique.

Visual Entailment Module For existing VG methods, fine-tuning with limited data hardly equips them with the ability to effectively handle irrelevant text and images. Hence, addressing the weak correlation between images and text in GMNER poses the primary challenge in integrating VG methods into GMNER. To achieve this objective, RiVEG introduces the VE module before the VG module.

For the classic VE task, denote the VE dataset $\mathcal{D}_{\text{VE}}$ as:

$$\mathcal{D}_{\text{VE}} = \{(i_1, h_1, l_1), \cdots, (i_n, h_n, l_n)\}$$

where i_i, h_i, l_i represent an image premise, a text hypothesis and a class label, respectively. Three labels e, n or c are determined by the relationship conveyed by (i_i, h_i). Specifically, e (entailment) represents $i_i \vDash h_i$, n (neutral) represents $i_i \nvDash h_i \wedge i_i \nvDash \neg h_i$, c (contradiction) represents $i_i \vDash \neg h_i$. Similar but distinct, RiVEG defines the GMNER dataset $\mathcal{D}_{\text{GMNER(VE)}}$ as:

$$\mathcal{D}_{\text{GMNER(VE)}} = \{(i_1, h_1, l_1), \cdots, (i_m, h_m, l_m)\}$$

where i_i, h_i, l_i represent an image that corresponds to a named entity, a named entity referring expression and a class label, respectively. Note that since the groundability of a named entity in GMNER dataset is unambiguous, only two labels e and c are retained. e represents h_i is groundable and c represents h_i is ungroundable. Finally, RiVEG replaces the original VE dataset with $\mathcal{D}_{\text{GMNER(VE)}}$ to fine-tune the existing vanilla VE model. In the inference phase, only the samples deemed groundable by the fine-tuned vanilla VE model will be fed into the final VG module. For the remaining samples that are judged to be ungroundable, their entity grounding results are directly defined as $None$.

Visual Grounding Module Expansion expressions bring named entities closer to the referring expressions required by existing VG methods. And VE module effectively filters out ungroundable samples. Once there is no requirement for VG methods to assess the groundability of the text input, existing vanilla VG methods can be seamlessly applied to determine the visually grounded region of the named entity. Denote the subset of all groundable named entity samples in GMNER dataset as:

$$\mathcal{D}_{\text{GMNER(VG)}} = \{(i_1, h_1, v_1), \cdots, (i_k, h_k, v_k)\}$$

where i_i, h_i, v_i represent an image that corresponds to a named entity, a named entity referring expression and a visually grounded region of h_i, respectively. Note that if a named entity in $\mathcal{D}_{\text{GMNER(VG)}}$ corresponds to multiple different bounding boxes, we specify that the bounding box with the largest area is chosen as the unique gold label. Subsequently, RiVEG employs $\mathcal{D}_{\text{GMNER(VG)}}$ to substitute the original VG dataset and fine-tune the existing vanilla VG model.

Methods	GMNER			MNER			EEG		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Text									
HBiLSTM-CRF-None (Lu et al., 2018a)	43.56	40.69	42.07	78.80	72.61	75.58	49.17	45.92	47.49
BERT-None (Devlin et al., 2019)	42.18	43.76	42.96	77.26	77.41	77.30	46.76	48.52	47.63
BERT-CRF-None (Yu et al., 2023)	42.73	44.88	43.78	77.23	78.64	77.93	46.92	49.28	48.07
BARTNER-None (Yan et al., 2021)	44.61	45.04	44.82	79.67	79.98	79.83	48.77	49.23	48.99
Text+Image									
GVATT-RCNN-EVG (Lu et al., 2018a)	49.36	47.80	48.57	78.21	74.39	76.26	54.19	52.48	53.32
UMT-RCNN-EVG (Yu et al., 2020)	49.16	51.48	50.29	77.89	79.28	78.58	53.55	56.08	54.78
UMT-VinVL-EVG (Yu et al., 2020)	50.15	52.52	51.31	77.89	79.28	78.58	54.35	56.91	55.60
UMGF-VinVL-EVG (Zhang et al., 2021a)	51.62	51.72	51.67	79.02	78.64	78.83	55.68	55.80	55.74
ITA-VinVL-EVG (Wang et al., 2022c)	52.37	50.77	51.56	80.40	78.37	79.37	56.57	54.84	55.69
BARTMNER-VinVL-EVG (Yu et al., 2023)	52.47	52.43	52.45	80.65	80.14	80.39	55.68	55.63	55.66
H-Index (Yu et al., 2023)	56.16	56.67	56.41	79.37	80.10	79.73	60.90	61.46	61.18
RiVEG$_{\text{BERT}}$ (ours)	64.03	63.57	63.80	83.12	82.66	82.89	67.16	66.68	66.92
RiVEG$_{\text{XLMR}}$ (ours)	65.09	66.68	65.88	84.80	84.23	84.51	67.97	69.63	68.79
△	+8.93	+10.01	+9.47	+5.43	+4.13	+4.78	+7.07	+8.17	+7.61
RiVEG$_{\text{BERT}}$† (ours)	65.10	66.96	66.02	83.02	85.40	84.19	68.21	**70.18**	69.18
RiVEG$_{\text{XLMR}}$† (ours)	**67.02**	**67.10**	**67.06**	**85.71**	**86.16**	**85.94**	**69.97**	70.06	**70.01**
△†	+10.86	+10.43	+10.65	+6.34	+6.06	+6.21	+9.07	+8.6	+8.83

Table 2: Performance comparison on the Twitter-GMNER dataset. All baseline results are from Yu et al. (2023). The model marked with † utilizes the data augmentation method. Detailed evaluation metrics are provided in Appendix A.4. And the model configuration details of RiVEG are detailed in Appendix A.1. △ and △† indicate the performance improvement compared with the previous state-of-the-art method H-Index.

4 Experiments

4.1 Settings

Datasets Based on two benchmark MNER datasets, *i.e.*, Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Yu et al., 2020), Yu et al. (2023) builds the Twitter-GMNER dataset by filtering samples with missing images or more than 3 entities belonging to the same type. More details about the training data for different modules and the Twitter-GMNER dataset are provided in Appendix A.2.

4.2 Main Results

We compare RiVEG with all existing baseline methods in Table 2. Following Yu et al. (2023), we also report results on two subtasks of GMNER, including MNER and Entity Extraction & Grounding (EEG). MNER aims to identify entity-type pairs, while EEG aims to extract entity-region pairs. The first group of text-only methods focuses on extracting entity-type pairs from text inputs. And the region prediction is then specified as the majority class, *i.e., None*. The second group of multimodal methods comprises the EVG series baseline methods and the end-to-end H-Index method proposed by Yu et al. (2023).

The experimental results demonstrate the superiority of RiVEG over previous methods. Compared with the previous state-of-the-art method H-Index, RiVEG exhibits significant advantages, achieving absolute improvements of 9.47%, 4.78% and 7.61% in all three tasks without any data augmentation. Additionally, data augmentation methods exhibit significant effects. After data augmentation, RiVEG achieves an additional 1.18%, 1.43%, and 1.22% performance improvement, respectively. This illustrates the rationality and effectiveness of our motivation to maximize the MNER effect and introduce the VE module and entity expansion expressions to use advanced VG methods to solve the GMNER task. We attribute the performance improvement of RiVEG to the following factors: 1) Maximizing the effectiveness of MNER significantly mitigates the issue of error propagation in GMNER predictions. Because if there is a prediction error in the entity or entity type within an entity-type-region prediction triplet, the triplet must be judged as an error. RiVEG alleviates this problem by making the MNER stage independent and introducing auxiliary knowledge. 2) RiVEG does not utilize the OD method for pre-extracting features from candidate regions. As shown in Table

MNER	Twitter-2015			Twitter-2017		
	Pre.	Rec.	F1	Pre.	Rec.	F1
Baseline	76.45	78.22	77.32	88.46	90.23	89.34
ITA(OCR+OD+IC)	-	-	78.03	-	-	89.75
$\text{MoRe}_{\text{Text}}$(Wiki)	-	-	77.91	-	-	89.50
$\text{MoRe}_{\text{Image}}$(Wiki)	-	-	78.13	-	-	89.82
LlaMA2-7B	78.16	79.57	78.86	90.12	91.19	90.66
LlaMA2-13B	78.37	**79.86**	79.10	89.57	92.15	90.84
Vicuna-7B	76.97	79.42	78.17	89.35	91.27	90.30
Vicuna-13B	78.46	78.92	78.69	90.16	90.23	90.20
PGIM(ChatGPT)	**79.21**	79.45	79.33	90.86	92.01	91.43
RiVEG(Ours)	79.10	79.78	**79.44**	**91.12**	**92.67**	**91.89**
	±0.46	±0.22	±0.17	±0.23	±0.09	±0.08

Table 3: The effectiveness of auxiliary knowledge provided by different LLMs in the MNER stage. The text encoder used by all methods is XLM-RoBERTa$_{\text{large}}$. Baseline means no auxiliary knowledge is added.

LLMs in GMNER_{VE}(Acc.)	**ALBEF**	**$\text{OFA}_{\text{large(VE)}}$**
Baseline	80.39	82.08
Vicuna-7B (Chiang et al., 2023)	82.13	82.57
Vicuna-13B (Chiang et al., 2023)	82.48	83.16
LlaMA2-7B (Touvron et al., 2023)	82.09	83.27
LlaMA2-13B (Touvron et al., 2023)	82.62	83.75
ChatGPT (Ouyang et al., 2022)	**83.36**	**84.30**
Mix	82.65	84.12

LLMs in GMNER_{VG}(Prec@0.5)	**SeqTR**	**$\text{OFA}_{\text{large(VG)}}$**
Baseline	67.27	72.40
Vicuna-7B (Chiang et al., 2023)	71.43	73.05
Vicuna-13B (Chiang et al., 2023)	71.51	73.26
LlaMA2-7B (Touvron et al., 2023)	70.98	72.89
LlaMA2-13B (Touvron et al., 2023)	71.39	73.45
ChatGPT (Ouyang et al., 2022)	72.10	73.90
Mix	**73.06**	**74.50**

Table 4: Results of the same model on datasets built by different LLMs. Baseline uses original named entities and their entity types. The rest of tests use their respective generated named entity referring expression.

1, existing OD methods exhibit subpar performance on the Twitter-GMNER dataset. RiVEG leverages the feature of VG methods, eliminating the need for pre-extracting features and naturally addressing this deficiency. 3) RiVEG designs an independent VE module to specifically tackle the weak correlation between image-text pairs in GMNER task. This method of enabling the model to undergo targeted training based on data characteristics enables RiVEG to gain a better ability to determine the groundability of named entities.

Note that this result does not represent the performance upper limit of RiVEG. The modular design facilitates straightforward upgrades, selecting more advanced methods for different modules in the future is expected to yield improved performance.[4]

4.3 Detailed Analysis

Contributions of Different LLMs to Various Subtasks Table 3 shows the test set results after using the same MNER model trained on different LLM versions of the same dataset. The method of obtaining external knowledge is exactly the same as PGIM (Li et al., 2023). Compared with ITA (Wang et al., 2022c) (Using optical character recognition, object detection and image caption) and MoRe (Wang et al., 2022b) (Using Wikipedia), the auxiliary knowledge provided by LLMs is more helpful to MNER. The quality of knowledge provided by LLMs with a small number of parameters is usually inferior to ChatGPT used by PGIM. RiVEG performs better than past state-of-the-art methods.[5] The training set of RiVEG includes responses from all five LLMs. And the dev set and test set are the same as PGIM. This indicates that utilizing diverse styles of external knowledge based on the same original sample enhances the generalization ability of model.

The named entities and their visual grounding results in the Twitter-GMNER dataset are deterministic, so using different LLMs to generate different named entity expansion expressions can build different LLM versions of the dataset. Table 4 shows the test set results of the same VE and VG models after training on different versions of the datasets. The training set of the Mix version is a direct merger of all five training sets, and the dev set is identical to the ChatGPT version. Using all versions of the named entity referring expression as text input performs significantly better than using raw named entities. LLMs with a large number of parameters generally generate better quality results than LLMs with a small number of parameters. The results of the Mix version show that more different styles of generated data can improve model performance. This is reflected more intuitively in Appendix A.6.

[4] Furthermore, we test 10 other text and visual variations of RiVEG in detail in Appendix A.5 and A.6. Experimental results show that the performance of all 14 RiVEG variants is superior to the previous state-of-the-art method.

[5] This is average result from three different random seeds.

VE Module Text Input (Acc.)	**ALBEF**	**OFA$_{\text{large(VE)}}$**
Named Entity+Entity Category	80.39	82.08
Entity Expansion Expression	81.76	81.40
Named Entity Referring Expression	**83.36**	**84.30**
VG Module Text Input (Prec@0.5)	**SeqTR**	**OFA$_{\text{large(VG)}}$**
Named Entity+Entity Category	67.27	72.40
Entity Expansion Expression	71.72	71.10
Named Entity Referring Expression	**72.10**	**73.90**

Table 5: The influence of varying text inputs on the performance of different methods.

VE Training Data	OFA$_{\text{base(VE)}}$	OFA$_{\text{large(VE)}}$
Pretraining+SNLI-VE+$\mathcal{D}_{\text{GMNER(VE)}}$	82.90	83.32
Pretraining+$\mathcal{D}_{\text{GMNER(VE)}}$	**83.41**	**84.30**
VG Training Data	OFA$_{\text{base(VG)}}$	OFA$_{\text{large(VG)}}$
Pretraining+RefCOCOg+$\mathcal{D}_{\text{GMNER(VG)}}$	71.60	73.64
Pretraining+RefCOCO++$\mathcal{D}_{\text{GMNER(VG)}}$	71.91	73.50
Pretraining+RefCOCO+$\mathcal{D}_{\text{GMNER(VG)}}$	**72.40**	72.52
Pretraining+$\mathcal{D}_{\text{GMNER(VG)}}$	68.83	**73.90**

Table 6: Performance after fine-tuning with various versions of pretrained weights on different datasets.

Named Entity Referring Expressions Analysis We further investigate the performance of various methods in VE and VG modules when confronted with diverse text inputs. The named entity referring expression in $\mathcal{D}_{\text{GMNER(VE/VG)}}$ is segmented into the named entity along with entity category and the entity expansion expression. Table 5 shows the test set performance of different methods after fine-tuning on $\mathcal{D}_{\text{GMNER(VE/VG)}}$. In order to control variables, this experiment is conducted on the expansion expression generated by ChatGPT.

The experimental results demonstrate that when the text inputs are the complete named entity referring expressions, all methods achieve optimal results. When the text inputs are named entities or entity expansion expressions, the performance of all methods decreases to a certain extent. Moreover, the sensitivity of different methods to various text inputs varies significantly. Unified frameworks like OFA demonstrate significantly better performance on fine-grained named entities than on coarse-grained entity expansion expressions. Contrarily, specialized methods like SeqTR, designed for a specific task, exhibit the opposite behavior. This experiment verifies the rationality of our proposed named entity referring expression. While different methods exhibit distinct characteristics, the expression method designed by us that combines both coarse and fine granularity emerges as their optimal choice.

Effect Analysis of Additional Training Data As the $\mathcal{D}_{\text{GMNER(VE/VG)}}$ datasets we create differ only in text type from the SNLI-VE (Xie et al., 2019) and RefCOCO series datasets (Mao et al., 2016; Yu et al., 2016; Nagaraja et al., 2016), we investigate whether pretraining with more similar task data could yield additional performance improvements in Table 6.

For various VE methods, incorporating additional fine-tuning on the SNLI-VE dataset (Xie et al., 2019) prior to the final fine-tuning on the $\mathcal{D}_{\text{GMNER(VE)}}$ does not result in improved performance. This observation suggests a significant difference between the $\mathcal{D}_{\text{GMNER(VE)}}$ and the SNLI-VE. For various VG methods, conducting additional fine-tuning on the RefCOCO series of datasets (Mao et al., 2016; Yu et al., 2016; Nagaraja et al., 2016) prior to the final fine-tuning on the $\mathcal{D}_{\text{GMNER(VG)}}$ may result in enhanced performance. This may be attributed to a certain degree of similarity between the two. Because we convert named entities into named entity referring expressions through entity expansion expressions.[6]

5 Conclusion

In this paper, we propose RiVEG, a unified framework that aims to use LLMs as bridges to reformulate the GMNER task into a unified MNER-VE-VG task. This reformulation enables the framework to seamlessly incorporate the capabilities of state-of-the-art methods for each subtask in their corresponding modules. Extensive experiments show that RiVEG significantly outperforms state-of-the-art methods. The proposed RiVEG greatly reduces the constraints of VG and VE methods on natural language inputs, expands their applicability to more realistic scenarios, and provides insights for future work to better deploy VG and VE modules in related tasks.

[6] ALBEF (Li et al., 2021) authors do not provide pretrained weights based on the SNLI-VE dataset. And some of the pretrained weights of SeqTR (Zhu et al., 2022) based on the RefCOCO series dataset cannot be obtained. Therefore, we follow the wishes of author and only test OFA(Wang et al., 2022a) in this experiment.

Limitations

In the past, the three-stage reformulation of GMNER was not promising by researchers. Because there may be two intuitive limitations to this reformulation. Limitation 1. Splitting the entire pipeline into three stages can lead to severe error propagation. Limitation 2. Pipeline design seems to be quite heavy. But actually, regarding Limitation 1, this paper demonstrates for the first time that the three-stage pipeline design performs significantly better than existing end-to-end methods. Regarding Limitation 2, the modular nature of RiVEG allows it to freely select lighter sub-models to form more efficient variants to deal with time-critical scenarios. The most lightweight variant of RiVEG can complete the entire three-stage training in 5 hours using a single 4090 GPU, making the training cost quite low. And the inference speed of RiVEG is also acceptable, since LLMs can be called in the form of API. Considering the substantial performance improvement, sacrificing a certain degree of inference speed is worthwhile. In the future, our research direction is to strive for optimal performance while designing each module in a lightweight style and training them in conjunction, possibly incorporating intermediate auxiliary losses.

Ethics Statement

In this paper, we use publicly available Twitter-GMNER dataset for experiments. For the auxiliary refined knowledge and the entity expansion expression, RiVEG generates them using ChatGPT, Vicuna, LlaMA2. Therefore, we trust that all the data we use does not violate the privacy of any user.

Acknowledgments

The authors would like to thank four anonymous reviewers for their insightful comments. This work was supported by the Natural Science Foundation of Tianjin (No.21JCYBJC00640) and the 2023 CCF-Baidu Songguo Foundation (Research on Scene Text Recognition Based on PaddlePaddle).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022a. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 904–915.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2024. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8032–8040.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023. Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2787–2802, Singapore. Association for Computational Linguistics.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018a. 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)*, pages 1990–1999.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018b. 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)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 817–834. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jieming Wang, Ziyan Li, Jianfei Yu, Li Yang, and Rui Xia. 2023. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3934–3943.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022b. Named entity and relation extraction with multi-modal retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5925–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022c. ITA: Image-text alignments for multimodal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States. Association for Computational Linguistics.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022d. Promptmner: prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *International Conference on Database Systems for Advanced Applications*, pages 297–305. Springer.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022e. Cat-mner: multimodal named entity recognition with knowledge-refined cross-modal attention. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. 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)*, pages 9141–9154, Toronto, Canada. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1114–1120.

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 *Proceedings of the AAAI conference on artificial intelligence*, 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

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

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer.

A Appendix

A.1 Model Configuration

In order to reasonably evaluate the capabilities of RiVEG, we choose advanced methods for each of three sub-modules of RiVEG. Specifically, for MNER module, RiVEG uses the same architecture as PGIM (Li et al., 2023), and selects BERT$_{\text{base}}$ (Devlin et al., 2019) and XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2020) as the text encoder, aligning with the prevailing MNER methods (Wang et al., 2022b,d,e). mPLUG-Owl (Ye et al., 2023) is employed to generate the image caption necessary for the entity expansion expression. For VE module, on the basis of OFA$_{\text{large}}$ (Wang et al., 2022a), RiVEG replaces its original VE dataset SNLI-VE (Xie et al., 2019) with $\mathcal{D}_{\text{GMNER(VE)}}$ and $\mathcal{D}_{\text{GMNER(VE)}^{\dagger}}$ for fine-tuning. For VG module, RiVEG chooses OFA$_{\text{large}}$ (Wang et al., 2022a), which also possesses visual grounding capability, to substitute its original VG datasets RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016; Nagaraja et al., 2016) with $\mathcal{D}_{\text{GMNER(VG)}}$ and $\mathcal{D}_{\text{GMNER(VG)}^{\dagger}}$ for fine-tuning.

A.2 Details of Different Module Datasets

Table 7 counts the training data used by different modules. And Table 8 shows the details of the Twitter-GMNER dataset. For the MNER module of RiVEG, the training data for base version comprises original samples and corresponding auxiliary refined knowledge generated by gpt-3.5-turbo provided by PGIM (Li et al.,

Split	**MNER Module Datasets**		**VE Module Datasets**		**VG Module Datasets**	
	#Twitter-GMNER	**#Twitter-GMNER**†	$\#\mathcal{D}_{\text{GMNER(VE)}}$	$\#\mathcal{D}_{\text{GMNER(VE)}^\dagger}$	$\#\mathcal{D}_{\text{GMNER(VG)}}$	$\#\mathcal{D}_{\text{GMNER(VG)}^\dagger}$
Train	7000	35000	11782	58910	4694	23470
Dev	1500	1500	2453	2453	986	986
Test	1500	1500	2543	2543	1036	1036
Total	10000	38000	16778	63906	6716	25492

Table 7: Statistics of datasets used in training of different sub-modules.

Split	**#Tweet**	**#Entity**	**#Groundable Entity**	**#Box**
Train	7000	11782	4694	5680
Dev	1500	2453	986	1166
Test	1500	2543	1036	1244
Total	10000	16778	6716	8090

Table 8: Statistics of Twitter-GMNER dataset (Yu et al., 2023).

2023). Additionally, data augmentation is applied to the training set of base version. In data augmentation version, we apply four additional LLMs (*i.e.*, vicuna-7b-v1.5, vicuna-13b-v1.5, llama-2-7b-chat-hf, llama-2-13b-chat-hf) to training set samples using the same algorithm, generating multiple versions of auxiliary refined knowledge for the same sample. The dev set and test set remain the same as base version to ensure the benchmarking of evaluation.

For VE module, the training data of base version $\mathcal{D}_{\text{GMNER(VE)}}$ exclusively comprises entity expansion expressions generated by gpt-3.5-turbo. In the data augmentation version $\mathcal{D}_{\text{GMNER(VE)}^\dagger}$, we also apply four additional LLMs to generate additional training samples and maintain the dev set and test set identical to base version $\mathcal{D}_{\text{GMNER(VE)}}$.

For VG module, the training data of base version $\mathcal{D}_{\text{GMNER(VG)}}$ is a subset of all groundable entity expansion expressions in base version $\mathcal{D}_{\text{GMNER(VE)}}$. Similarly, the data augmentation version $\mathcal{D}_{\text{GMNER(VG)}^\dagger}$ is a subset of the data augmentation version $\mathcal{D}_{\text{GMNER(VE)}^\dagger}$.

A.3 Detailed Implementation Details

For the methods used by three modules of RiVEG, we only replace their original VE and VG datasets with $\mathcal{D}_{\text{GMNER(VE)}}/\mathcal{D}_{\text{GMNER(VE)}^\dagger}$ and $\mathcal{D}_{\text{GMNER(VG)}}/\mathcal{D}_{\text{GMNER(VG)}^\dagger}$ and primarily adhere to original hyperparameter configurations and training strategies as reported in their paper for fine-tuning. Without specific instructions, all training is conducted using a single 4090 GPU. The model with the best results on each dev set is selected to evaluate the performance on each test set and used to perform the final inference. Note that considering the notable effectiveness of RiVEG, we refrain from conducting attempts with excessive hyperparameter combinations. Thus, exploring more hyperparameter combinations may yield improved results.

MNER Module RiVEG mainly follows MNER fine-tuning strategy of PGIM[7] (Li et al., 2023). For the XLM-RoBERTa$_{\text{large}}$ version in Table 2, Table 9 and Table 10, learning rate is set to 7e-6 and batchsize is set to 4. For the BERT$_{\text{base}}$ version in Table 10, learning rate is set to 5e-5 and batchsize is set to 16. Remaining settings remain unchanged. The model is fine-tuned for 25 epochs.

VE Module For OFA$_{\text{large(VE)}}$ (Wang et al., 2022a), the learning rate is set to 2e-5 and batchsize is set to 32. We convert the labels entailment/contradiction to yes/no. And we also use the Trie-based search strategy to constrain the generated labels over the candidate set. Remaining settings follow their VE fine-tuning strategy[8]. The basic version of OFA$_{\text{large}}$ is fine-tuned for 6 epochs on $\mathcal{D}_{\text{GMNER(VE)}}/\mathcal{D}_{\text{GMNER(VE)}^\dagger}$.

For ALBEF (Li et al., 2021), the learning rate is set to 2e-5 and batchsize is set to 24. Remaining settings follow their VE fine-tuning strategy[9]. We fine-tune for 5 epochs on $\mathcal{D}_{\text{GMNER(VE)}}/\mathcal{D}_{\text{GMNER(VE)}^\dagger}$ using a single V100 32G GPU based on ALBEF-14M.

VG Module For OFA$_{\text{large(VG)}}$ (Wang et al., 2022a), learning rate is 3e-5 with the label smoothing of 0.1 and batchsize is set to 32. The input image resolution is set to 512×512. And the basic version of OFA$_{\text{large}}$ is fine-tuned for 10 epochs

[7] https://github.com/JinYuanLi0012/PGIM

[8] https://github.com/OFA-Sys/OFA#visual-entailment

[9] https://github.com/salesforce/ALBEF#visual-entailment

Visual Variants of RiVEG	**GMNER**			**MNER**			**EEG**		
MNER(F1)+VE(Acc.)+VG(Prec@0.5)	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Baseline (H-Index (Yu et al., 2023))	56.16	56.67	56.41	79.37	80.10	79.73	60.90	61.46	61.18
w/o Data Augmentation									
PGIM(84.51)+$\text{OFA}_{\text{large(VE)}}$(84.30)+SeqTR(72.10)	57.06	58.46	57.75	84.80	84.23	84.51	59.79	61.25	60.51
PGIM(84.51)+ALBEF-14M(83.36)+SeqTR(72.10)	57.45	58.85	58.14	84.80	84.23	84.51	60.25	61.72	60.98
PGIM(84.51)+ALBEF-14M(83.36)+$\text{OFA}_{\text{large(VG)}}$(73.90)	64.56	66.13	65.33	84.80	84.23	84.51	67.47	69.12	68.29
PGIM(84.51)+$\text{OFA}_{\text{large(VE)}}$(84.30)+$\text{OFA}_{\text{large(VG)}}$(73.90)	65.09	66.68	65.88	84.80	84.23	84.51	67.97	69.63	68.79
Data Augmentation									
PGIM(85.94)+$\text{OFA}_{\text{large(VE)}}$(84.12)+SeqTR(73.06)	59.03	59.10	59.06	85.71	86.16	85.94	61.85	61.93	61.89
PGIM(85.94)+ALBEF-14M(82.65)+SeqTR(73.06)	59.14	59.22	59.18	85.71	86.16	85.94	62.01	62.09	62.05
PGIM(85.94)+ALBEF-14M(82.65)+$\text{OFA}_{\text{large(VG)}}$(74.50)	66.12	66.20	66.16	85.71	86.16	85.94	68.99	69.07	69.03
PGIM(85.94)+$\text{OFA}_{\text{large(VE)}}$(84.12)+$\text{OFA}_{\text{large(VG)}}$(74.50)	67.02	67.10	67.06	85.71	86.16	85.94	69.97	70.06	70.01

Table 9: Performance comparison of different visual variants of RiVEG. Here we keep MNER module unchanged to clearly illustrate the impact of different VE and VG methods on performance. Therefore, different variants of the same group have the same MNER results. The test results of more text variants are shown in Table 10.

on $\mathcal{D}_{\text{GMNER(VG)}}/\mathcal{D}_{\text{GMNER(VG)}^{\dagger}}$. Remaining settings follow their VG fine-tuning strategy[10].

For SeqTR (Zhu et al., 2022), since SeqTR does not provide pretrained weights, we select the weight which are pretrained and fine-tuned for 5 epochs on RefCOCO[11] (Yu et al., 2016), as the initial weight and fine-tune for 5 epochs on $\mathcal{D}_{\text{GMNER(VG)}}/\mathcal{D}_{\text{GMNER(VG)}^{\dagger}}$. The input image resolution is set to 640×640. Text length is trimmed to 30. The learning rate is 5e-4 and batchsize is 64. Visual encoder is YOLOv3 (Redmon and Farhadi, 2018). Remaining settings follow their Referring Expression Comprehension fine-tuning strategy[12].

A.4 Evaluation Metrics

The GMNER prediction is composed of entity, type, and visual region. Following Yu et al. (2023), the correctness of each prediction is computed as follows:

$$C_e/C_t = \begin{cases} 1, & p_e/p_t = g_e/g_t; \\ 0, & \text{otherwise.} \end{cases}$$

$$C_v = \begin{cases} 1, & p_v = g_v = None; \\ 1, & \max(IoU_1, \cdots, IoU_j) > 0.5; \\ 0, & \text{otherwise.} \end{cases}$$

$$correct = \begin{cases} 1, & C_e \wedge C_t \wedge C_v = 1; \\ 0, & \text{otherwise.} \end{cases}$$

where C_e, C_t and C_v represent the correctness of entity, type and region predictions; p_e, p_t and p_v represent the predicted entity, type and region; g_e, g_t and g_v represent the gold entity, type and region; and IoU_j denotes the IoU score between p_v with the j-th ground-truth bounding box $g_{v,j}$. Then precision (Pre.), recall (Rec.) and F1 score are used to evaluate the performance of model:

$$Pre = \frac{\#correct}{\#predict}, Rec = \frac{\#correct}{\#gold}$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}$$

where $\#correct$, $\#predict$ and $\#gold$ respectively represent the number of triples of correct predictions, predictions and gold labels.

A.5 Exploration of More Visual Variants

To further demonstrate the effectiveness of RiVEG, we present the performance of various RiVEG variants across all three subtasks in Table 9. Specifically, we additionally choose an early classic multimodal pretraining method ALBEF-14M (Li et al., 2021) for VE module and an early classic VG method SeqTR (Zhu et al., 2022) for VG module. Different combinations of methods are considered as different variants. Experimental results show that, without considering any data augmentation, the weakest variant still exhibits highly competitive results with the previous state-of-the-art method. And all the variants after data augmentation are significantly better than the baseline.

We also present the corresponding test set results for different methods of different modules

[10] https://github.com/OFA-Sys/OFA#visual-grounding-referring-expression-comprehension

[11] https://github.com/seanzhuh/SeqTR#refcoco

[12] https://github.com/seanzhuh/SeqTR#pre-training--fine-tuning

Text Variants of RiVEG MNER(F1)+VE(Acc.)+VG(Prec@0.5)	**GMNER**			**MNER**			**EEG**		
	Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Baseline									
Unimodal Baseline (BERT-CRF-None (Yu et al., 2023))‡	42.73	44.88	43.78	77.23	78.64	77.93	46.92	49.28	48.07
Multimodal Baseline (H-Index (Yu et al., 2023))‡	56.16	56.67	56.41	79.37	80.10	79.73	60.90	61.46	61.18
w/o Data Augmentation									
PGIM$_{\text{BERT}}$(82.89)+OFA$_{\text{large(VE)}}$(84.30)+OFA$_{\text{large(VG)}}$(73.90)	64.03	63.57	63.80	83.12	82.66	82.89	67.16	66.68	66.92
PGIM$_{\text{XLMR}}$(84.51)+OFA$_{\text{large(VE)}}$(84.30)+OFA$_{\text{large(VG)}}$(73.90)	65.09	66.68	65.88	84.80	84.23	84.51	67.97	69.63	68.79
Data Augmentation									
PGIM$_{\text{BERT}}$(84.19)+OFA$_{\text{large(VE)}}$(84.12)+OFA$_{\text{large(VG)}}$(74.50)	65.10	66.96	66.02	83.02	85.40	84.19	68.21	**70.18**	69.18
PGIM$_{\text{XLMR}}$(85.94)+OFA$_{\text{large(VE)}}$(84.12)+OFA$_{\text{large(VG)}}$(74.50)	**67.02**	**67.10**	**67.06**	**85.71**	**86.16**	**85.94**	**69.97**	70.06	**70.01**
w/o Auxiliary Refined Knowledge in MNER									
BERT-CRF(78.19)+OFA$_{\text{large(VE)}}$(84.30)+OFA$_{\text{large(VG)}}$(73.90)	61.16	59.95	60.55	78.93	77.47	78.19	65.33	64.04	64.68
BERT-CRF(78.19)+OFA$_{\text{large(VE)}}$(84.12)†+OFA$_{\text{large(VG)}}$(74.50)†	62.16	60.94	61.54	78.93	77.47	78.19	66.29	64.99	65.63
XLMR-CRF(82.90)+OFA$_{\text{large(VE)}}$(84.30)+OFA$_{\text{large(VG)}}$(73.90)	63.89	64.24	64.06	82.68	83.13	82.90	67.45	67.82	67.63
XLMR-CRF(82.90)+OFA$_{\text{large(VE)}}$(84.12)†+OFA$_{\text{large(VG)}}$(74.50)†	64.67	65.03	64.85	82.68	83.13	82.90	68.11	68.49	68.30

Table 10: Performance of different text variants of RiVEG. Different from Table 9, here we keep the VE and VG modules unchanged. For the baseline model, results of methods with ‡ come from Yu et al. (2023). † indicates that the VE and VG modules use data augmentation methods.

after fine-tuning. The results indicate that stronger combinations of sub-methods generally yield superior performance. The performance of the early classic ALBEF and SeqTR is inferior to the subsequent advanced OFA. Additionally, while the test set Prec@0.5 of SeqTR is slightly lower than that of OFA$_{\text{large(VG)}}$, there are significant differences in their final GMNER and EEG results. This phenomenon may be attributed to the text representation method employed by SeqTR. The GRU (Chung et al., 2014) vocabulary of SeqTR is solely statistically obtained from $\mathcal{D}_{\text{GMNER(VG)}}$ or $\mathcal{D}_{\text{GMNER(VG)}^{\dagger}}$, rendering it incapable of handling out-of-vocabulary words in new samples during inference.

Moreover, data augmentation methods exhibit varying effects on distinct modules. MNER and VG methods can intuitively benefit from data augmentation, but this is not the case for VE methods. One potential reason is that the distribution after fitting more training data does not align with the optimal distribution of the test data. But this is acceptable since data augmentation significantly enhances the performance of overall framework.

A.6 Exploration of More Text Variants

Table 10 explores the influence of various text encoders on the framework. Specifically, we fix the VE and VG modules and replace the originally used XLM-RoBERTa$_{\text{large}}$ in the MNER module with BERT$_{\text{base}}$. The main conclusions are as follows: 1) The weaker text encoder evidently results in inferior MNER performance compared with the XLM-RoBERTa variant. Due to the error propagation characteristics of GMNER prediction triples, weak MNER performance naturally results in the degradation of GMNER and EEG performance. 2) Concerning the MNER results, the BERT version of RiVEG outperforms the BERT version of Unimodal Baseline, surpassing 4.96% without any data augmentation and achieving a further lead of 6.26% after data augmentation. It underscores the effectiveness of incorporating auxiliary knowledge into the original samples in MNER task.

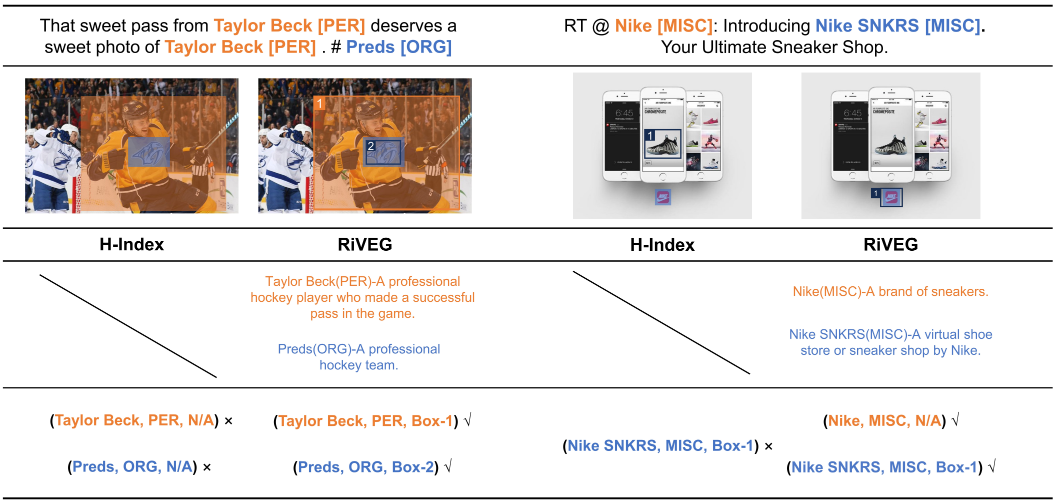
Additionally, to control variables and compare entity grounding performance across different methods, we conduct more challenging tests for RiVEG. The variants in the *w/o Auxiliary Refined Knowledge in MNER* group indicate that the text input to their MNER module consists only of the original text and does not incorporate any external knowledge. In this scenario, the MNER performance of these BERT variants of RiVEG is nearly identical to that of the two baseline methods. However, the GMNER and EEG performance of these variants still significantly outperforms all baseline methods. This comparison highlights the robust performance of RiVEG in entity grounding.

The last four lines of experimental results also illustrate the efficacy of the data augmentation method. When facing the same named entity recognition and expansion results, VE and VG models trained using data augmentation methods generally achieve better GMNER and EEG results. This echoes the experimental results in Table 4.

A.7 Case Study

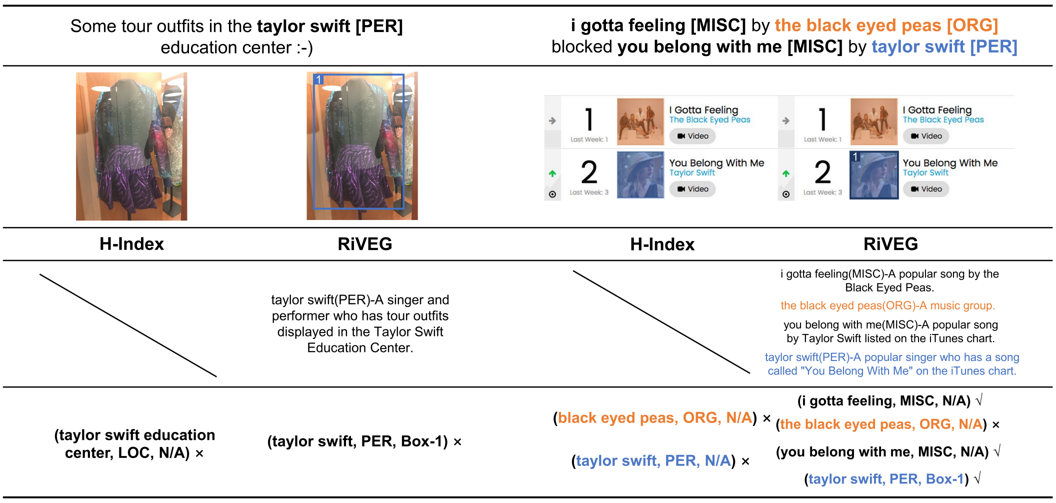
We further conduct case study to compare the predictions of H-Index (Yu et al., 2023) and RiVEG for challenging samples. Figure 3 illustrates two examples where H-Index makes errors, while RiVEG produces accurate predictions. For the test sample on the left, even though H-Index correctly predicts the entity-type pairs, it fails to achieve any effective entity grounding. All predictions made by RiVEG are accurate, even for the challenging entity grounding case of the "*Preds*" team logo. For the sample on the right, H-Index fails to predict all named entities and does not perform entity grounding correctly. RiVEG completes predictions perfectly, even though the grounded area of "*Nike SNKRS*" is very small.

Figure 4 illustrates two error examples. The left test sample poses a significant challenge for entity grounding due to the resemblance between the shape of the clothing in the image and that of humans. H-Index fails to predict the named entity. While RiVEG accurately predicts the entity-type pairs, it incorrectly performs entity grounding. The entity prediction and grounding for the sample on the right are both highly challenging. In this case, H-Index is almost incapable of making effective predictions. RiVEG accurately predicts all entity-type pairs and successfully performs the grounding of "*taylor swift*". However, it fails to accurately complete the grounding of "*the black eyed peas*". This indicates that the GMNER task remains highly challenging.



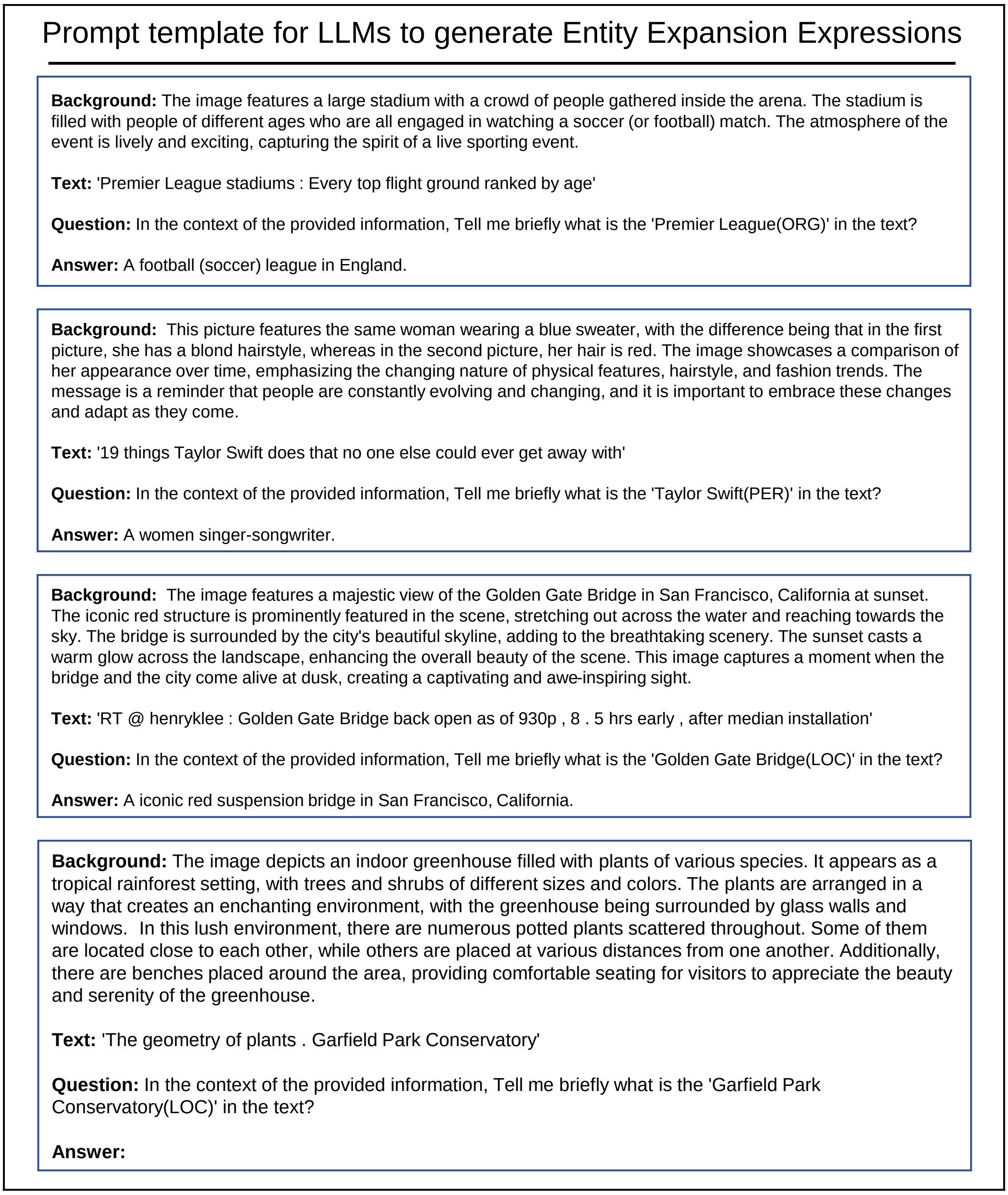
That sweet pass from **Taylor Beck [PER]** deserves a sweet photo of **Taylor Beck [PER]** . # **Preds [ORG]**		RT @ **Nike [MISC]**: Introducing **Nike SNKRS [MISC]**. Your Ultimate Sneaker Shop.	
H-Index	**RiVEG**	**H-Index**	**RiVEG**
	Taylor Beck(PER)-A professional hockey player who made a successful pass in the game. Preds(ORG)-A professional hockey team.		Nike(MISC)-A brand of sneakers. Nike SNKRS(MISC)-A virtual shoe store or sneaker shop by Nike.
(Taylor Beck, PER, N/A) × **(Preds, ORG, N/A)** ×	**(Taylor Beck, PER, Box-1)** √ **(Preds, ORG, Box-2)** √	**(Nike SNKRS, MISC, Box-1)** ×	**(Nike, MISC, N/A)** √ **(Nike SNKRS, MISC, Box-1)** √

Figure 3: Compare the predictions of RiVEG and H-Index for two test samples.



Some tour outfits in the **taylor swift [PER]** education center :-)		**i gotta feeling [MISC]** by **the black eyed peas [ORG]** blocked **you belong with me [MISC]** by **taylor swift [PER]**	
H-Index	**RiVEG**	**H-Index**	**RiVEG**
	taylor swift(PER)-A singer and performer who has tour outfits displayed in the Taylor Swift Education Center.		i gotta feeling(MISC)-A popular song by the Black Eyed Peas. the black eyed peas(ORG)-A music group. you belong with me(MISC)-A popular song by Taylor Swift listed on the iTunes chart. taylor swift(PER)-A popular singer who has a song called "You Belong With Me" on the iTunes chart.
(taylor swift education center, LOC, N/A) ×	**(taylor swift, PER, Box-1)** ×	**(black eyed peas, ORG, N/A)** × **(taylor swift, PER, N/A)** ×	**(i gotta feeling, MISC, N/A)** √ **(the black eyed peas, ORG, N/A)** × **(you belong with me, MISC, N/A)** √ **(taylor swift, PER, Box-1)** √

Figure 4: A case study on two incorrect predictions.



Prompt template for LLMs to generate Entity Expansion Expressions

Background: The image features a large stadium with a crowd of people gathered inside the arena. The stadium is filled with people of different ages who are all engaged in watching a soccer (or football) match. The atmosphere of the event is lively and exciting, capturing the spirit of a live sporting event.

Text: 'Premier League stadiums : Every top flight ground ranked by age'

Question: In the context of the provided information, Tell me briefly what is the 'Premier League(ORG)' in the text?

Answer: A football (soccer) league in England.

Background: This picture features the same woman wearing a blue sweater, with the difference being that in the first picture, she has a blond hairstyle, whereas in the second picture, her hair is red. The image showcases a comparison of her appearance over time, emphasizing the changing nature of physical features, hairstyle, and fashion trends. The message is a reminder that people are constantly evolving and changing, and it is important to embrace these changes and adapt as they come.

Text: '19 things Taylor Swift does that no one else could ever get away with'

Question: In the context of the provided information, Tell me briefly what is the 'Taylor Swift(PER)' in the text?

Answer: A women singer-songwriter.

Background: The image features a majestic view of the Golden Gate Bridge in San Francisco, California at sunset. The iconic red structure is prominently featured in the scene, stretching out across the water and reaching towards the sky. The bridge is surrounded by the city's beautiful skyline, adding to the breathtaking scenery. The sunset casts a warm glow across the landscape, enhancing the overall beauty of the scene. This image captures a moment when the bridge and the city come alive at dusk, creating a captivating and awe-inspiring sight.

Text: 'RT @ henryklee : Golden Gate Bridge back open as of 930p , 8 . 5 hrs early , after median installation'

Question: In the context of the provided information, Tell me briefly what is the 'Golden Gate Bridge(LOC)' in the text?

Answer: A iconic red suspension bridge in San Francisco, California.

Background: The image depicts an indoor greenhouse filled with plants of various species. It appears as a tropical rainforest setting, with trees and shrubs of different sizes and colors. The plants are arranged in a way that creates an enchanting environment, with the greenhouse being surrounded by glass walls and windows. In this lush environment, there are numerous potted plants scattered throughout. Some of them are located close to each other, while others are placed at various distances from one another. Additionally, there are benches placed around the area, providing comfortable seating for visitors to appreciate the beauty and serenity of the greenhouse.

Text: 'The geometry of plants . Garfield Park Conservatory'

Question: In the context of the provided information, Tell me briefly what is the 'Garfield Park Conservatory(LOC)' in the text?

Answer:

Figure 5: A prompt template for LLMs to generate entity expansion expressions.