

Multi-View Prompt for Fine-Grained Multimodal Named Entity Recognition and Grounding

Jintao Liu^{a,†}, Chenglong Liu^{a,†} and Kaiwen Wei^{b,*}

^aUniversity of Chinese Academy of Sciences, Beijing, China

^bCollege of Computer Science, Chongqing University, Chongqing, China

Abstract. Fine-Grained Multimodal Named Entity Recognition and Grounding (FMNERG) aims to extract entity name, fine-grained entity type, and its corresponding object from paired text and image. This task demands fundamental reasoning capability for complex language and multimodal comprehension. Despite encouraging results, existing methods face two critical issues: (1) Insufficient knowledge of the entity poses challenges to fine-grained entity recognition; (2) Limited correlations between entities and objects hinder the visual grounding of entities. To tackle these issues, we propose a Multi-View Prompt (MVP) method for the FMNERG task in this paper, which collaborates with Large Language Models (LLMs) and Visual Grounding Models (VGMs) for reasoning. Concretely, MVP constructs a knowledgeable prompt in a chain-of-thought format, progressively refining possible entity types from coarse-grained to fine-grained levels. It leverages a heuristic method to select demonstration examples, which could provide guiding knowledge about entities from LLMs. To establish correlations between entities and potential objects, MVP introduces a grounded prompt that exploits information from guiding knowledge and image caption, enabling VGMs to detect related objects. Experimental results indicate that MVP achieves state-of-the-art performance on the Twitter dataset.

1 Introduction

Multimodal Named Entity Recognition (MNER) aims to detect entity spans and classify them to corresponding entity types from the given text-image pair. Existing MNER studies mainly regard visual modality as supplementary information [34] and only four coarse-grained types are defined (*person*, *location*, *organization*, and *miscellaneous*) [25], limiting its application in downstream tasks. To bridge the research gap, FMNERG has been proposed to ground entities to objects in the image and classify entities into more fine-grained types. As shown in Fig. 1, given the image-text pair, an FMNERG system should be able to extract the entity *Kevin Durant* as *athlete* and detect the corresponding object in the image, and extract the other two triples in the same way. This task has a wide application in multimodal knowledge graph construction [17], multimodal entity disambiguation [19], visual question answer [1], and so on.

Previous methods either employ better representations of images and objects [8, 6] or use image captions [5, 28] to facilitate entity extraction. These methods typically regard MNER and entity grounding as two separate tasks, which might suffer from error propaga-

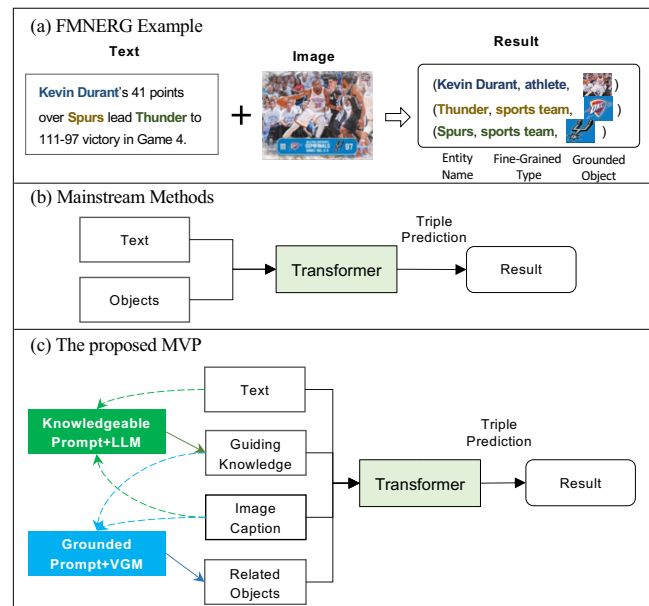


Figure 1. Illustration of an example of FMNERG (a) and comparison of using mainstream methods (b) and the proposed MVP method (c) to perform FMNERG task.

tion. Yu et al. [34] and Wang et al. [25] employ a generative method to simultaneously extract entity names, entity types, and their corresponding visual objects in the image. Despite the advancements achieved by these methods, they are not effective due to the following two critical issues:

(1) **Insufficient knowledge** about entities poses struggles to fine-grained entity recognition. For example, in Fig. 1, discerning *Thunder* as fine-grained type *sports team* tends to be more challenging than identifying it as coarse-grained type *organization*. Some studies [27, 39, 26] proposed the incorporation of external context from databases to improve the reasoning capability of the MNER model. But such knowledge might show inconsistencies from current text and introduce irrelevant information. Shao et al. [24] and Li et al. [11] adopted in-context learning method to prompt ChatGPT to generate knowledge. However, these models only encompass a degree of coarse-grained entity type knowledge, predicting fine-grained entity types remains an intricate task.

(2) **Restricted correlations** between entities in text and objects in image hinder the visual grounding of entities. Existing methods

* Corresponding Author. Email: weikaiwen@cqu.edu.cn

† Equal contribution.

neglect the utilization of related information about entities to guide the detection of candidate objects, and instead adopt general object detectors [36] to detect. These detectors are typically trained on datasets with predefined object types that differ from fine-grained entity types, resulting in a large number of unrelated objects and potentially a failure to cover ground truth boxes. Recently, open-set object detection methods represented by Grounding Dino [16] have demonstrated remarkable performance, which can detect corresponding objects based on categories in text. However, due to the sparsity of entity types mentioned in the text, directly introducing it would lead to serious missed detections.

In this paper, we propose a Multi-View Prompt (MVP) method for the FMNERG task, which collaborates with LLMs and VGMs via knowledgeable prompt and grounded prompt, respectively. Specifically, to cope with the first issue, we employ a knowledgeable prompt in a chain-of-thought format to refine entity type from coarse-grained to fine-grained levels, and select demonstration examples heuristically to encourage LLMs to generate guiding knowledge. To address the second issue, we introduce a grounded prompt consisting of nouns from image caption, fine-grained type from guiding knowledge, and pre-defined coarse-grained type to assist VGMs in detecting related objects, which can model potential correlations between entity and its object in the image. The text, guiding knowledge, image caption, and candidate objects are combined as input sequence. We leverage a pre-trained transformer-based architecture to generate the output sequence, where we can decode the entity-type-object triples. Experimental results illustrate that MVP achieves state-of-the-art performance on the benchmark dataset.

The main contributions of this paper can be summarised as follows:

- This paper proposes a MVP method for the FMNERG task, which collaborates with LLMs and VGMs to enhance the reasoning capability of the model.
- The knowledgeable prompt in MVP can guide the reasoning process of LLMs and benefit fine-grained entity recognition. The grounded prompt can assist VGMs in encompassing related objects and facilitate the capture of correlations between entities and objects.
- Experimental results indicate that the proposed method outperforms the state-of-the-art models, showing significant improvements on the benchmark dataset.

2 Related Work

2.1 Multimodal Named Entity Recognition

With the rapid development of multimodal posts on social media, the MNER task has emerged as a significant research area. Some approaches [40, 35, 15, 13] utilize visual modalities as auxiliary cues to enhance named entity recognition. Jia et al. [8] and Chen et al. [6] seek to obtain better visual representations, while [5, 28] use image caption or OCR text as visual context for image-text alignment. Considering the MNER outputs struggle in multimodal knowledge graph construction and entity disambiguation, Yu et al. [34] has advanced MNER by grounding entities to objects within the image and extracted the entity-type-region triples in a sequence-to-sequence manner. Wang et al. [25] further extends entity type to a fine-grained level and employs a generative method to simultaneously extract named entities, fine-grained entity types, and their corresponding objects in the image. However, the insufficient knowledge of the entity remains

a challenge, hindering the capabilities of fine-grained entity recognition.

2.2 Visual Grounding

Visual Grounding (VG) aims to locate the most relevant object or region in an image based on the natural language query, which can be divided into one-stage and two-stage methods. The first branch uses end-to-end object detection methods such as YOLO [20] and DETR [3], and fuse extra features to directly predict the regions [32, 7, 14, 12]. But these techniques pose major optimization challenges when combined with MNER. The second branch first obtains region proposals as candidate objects via object detection methods [22] and then ranks them based on the region-query relevance [31, 4]. Yu et al. [34] and Wang et al. [25] adopt VinVL [36] to extract candidate objects for entity visual grounding. However, these methods would generate a large number of irrelevant objects that may not fully encompass ground truth regions.

2.3 Prompting Paradigms

Several prompting paradigms have been proposed to enhance the reasoning abilities of LLMs without the need for model parameter updating, which can be divided into In-Context Learning (ICL) and chain-of-thought (CoT) methods. The ICL paradigm can learn to comprehend the task from the given demonstration examples [2]. Since the examples have a strong effect on the ICL performance, some studies focus on how to select effective examples, such as similarity-based retrieval method [23], gradient-based method [29]. The CoT prompting method aims to enhance the reasoning ability of LLMs by a series of intermediate inference steps [30], rather than directly providing the final answer. Kojima et al. [9] seeks to generate reasoning chains automatically via a simple prompt like *Let's think step by step*. In our work, we leverage the CoT prompt to query ChatGPT from coarse-grained to fine-grained levels to generate guiding knowledge and heuristically choose demonstration samples.

3 Preliminaries

Given an input sample including a text T and an accompanying image I , the objective of the Fine-grained Multimodal Named Entity Recognition and Grounding (FMNERG) task is to extract all entity-type-object triples from it:

$$E = \{(e_1, f_1, o_1), \dots, (e_m, f_m, o_m)\} \quad (1)$$

where (e_i, f_i, o_i) indicates the i -th triple, e_i denotes the entity which is a span of text, f_i denotes the fine-grained entity type of e_i , and o_i denotes the corresponding object of e_i in the image. In the Twitter dataset, there are 8 coarse-grained entity types, such as *person*, *building*, *organization*, etc. Each coarse-grained entity type contains multiple fine-grained entity types, for example, the *person* contains fine-grained type *athlete*, *actor*, and *musician*, etc. Note that each entity has a corresponding fine-grained entity type but might not have a grounded object. If e_i is grounded in the image, o_i is a 4-dimensional vector indicating the position of the corresponding object. In contrast, if e_i is not grounded in the image, o_i is set to *None*.

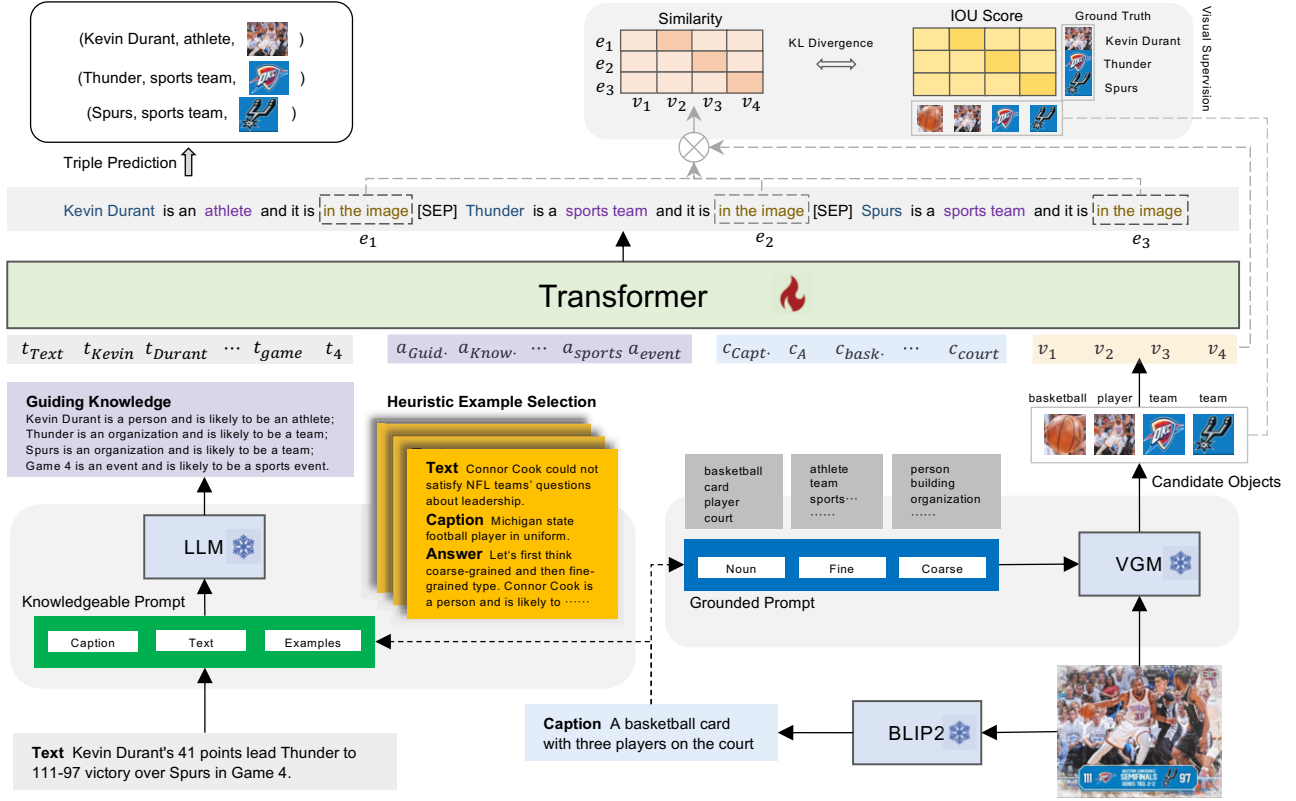


Figure 2. The overview framework of the proposed MVP method. The input for Transformer consists of text, guiding knowledge, image caption, and candidate objects.

4 Methodology

4.1 Overview

The overview of the proposed MVP method is illustrated in Fig. 2. We formulate FMNERG as a sequence-to-sequence generative problem. MVP leverages knowledgeable prompt and grounded prompt to generate guiding knowledge and related candidate objects, respectively. The knowledgeable prompt adopts a chain-of-thought mechanism to encourage the reasoning process of LLMs, and uses a heuristic approach to select demonstration examples for effective few-shot learning. The grounded prompt aims to cover more related objects and mine the subtle clues between entities and objects. Subsequently, we combine the text, guiding knowledge, image caption, and candidate objects as input for the Transformer model and extract entity-type-object triples from the output sequence.

4.2 Task Formulation

Previous MNER methods mainly adopt a classification model with a BIO tagging schema. However, this mechanism is hard to jointly optimize with entity visual grounding and is also difficult to adapt to FMNERG. In this paper, we solve the FMNERG task in an end-to-end generative manner, which can take the text and image as input and decode entity-type-object triples from the output.

We first construct the target output sequence for each triple (e_i, f_i, o_i) . The target output sequence can be formalized as:

$$e_i \text{ is a } f_i \text{ and it is (not) in the image} \quad (2)$$

where *not* in this sequence depends on whether e_i is grounded in the image. For example, in Fig. 2, the target output sequence of the first triple is *Kevin Durant is an athlete and it is in the image*, where *Kevin Durant* denotes entity name, *athlete* represent fine-grained entity type, and this entity is grounded in the image. Similarly, we can obtain the target sequences of the other two triples. Then we concatenate all these sequences with special token *[SEP]* to form the final target output sequence Y .

4.3 Knowledgeable Prompt

Our objective is to query LLMs to generate guiding knowledge for reasoning. Considering that ChatGPT API only accepts text modality as its input, in order to make image information understandable, we utilize an advanced multimodal model BLIP-2 [10] to transform image I into its caption P .

Chain-of-Thought. Due to the lack of knowledge about fine-grained entity types, it is difficult to predict them directly. To this end, this work adopts the CoT paradigm by taking easy-to-predict coarse-grained types as an intermediate reasoning process and then predicting possible fine-grained types. We can add the prompt words *Let's first think coarse-grained and then fine-grained type* before each answer to facilitate the reasoning process. Consequently, the input for ChatGPT includes text, image caption, and CoT prompt words (see Table 1).

Heuristic Example Selection. Since the demonstration examples exhibit a substantial influence on the few-shot learning ability of

Table 1. An example of the prompt template for zero-shot chain-of-thought reasoning.

Chain-of-thought Prompt
<p>Here are some content that people post on Twitter, and these content are composed of text and image caption. Notice: entity name exists only in ‘Text’, not in ‘Caption’, don’t change the writing style and format of entity names. The coarse-grained type include: person, organization, location, event, art, product, building, other. Note that if the Text has entity, the answer must be in form of: “[entity name] is a [coarse-grained type] and is likely to be a [fine-grained type]”.</p> <p>Text: # BPLStorySoFar Nemanja Matic of @ ChelseaFC has won the most tackles in the # BPL 2014/15 to date . . . Caption: the top 10 players in the premier league Question: Analyze the Text and the Caption, which named entities and their corresponding types are included in the Text? Answer: Let’s first think coarse-grained type and then fine-grained type step by step.</p>

Table 2. An example of the the construction of knowledgeable prompt and grounded prompt.

Knowledgeable Prompt
<p>Here are some content that people post on Twitter, and these content are composed of text and image caption. Notice: entity name exists only in ‘Text’, not in ‘Caption’, don’t change the writing style and format of entity names. The coarse-grained type include: person, organization, location, event, art, product, building, other. Note that if the Text has entity, the answer must be in form of: “[entity name] is a [coarse-grained type] and is likely to be a [fine-grained type]”.</p> <p>Text: Blackhawks vs . Sharks at the United Center . # NHL Caption: the chicago blackhawks are playing in an ice hockey game Question: Analyze the Text and the Caption, which named entities and their corresponding types are included in the Text? Answer: 1. Blackhawks is an organization and is likely to be a sports team 2. Sharks is an organization and is likely to be a sports team 3. United Center is a building and is likely to be a sports facility 4. NHL is an organization and is likely to be a sports league ...</p> <p>Text: RT @ thehill : ObamaCare win turns up heat on GOP presidential field Caption: four different pictures of men in suits and ties Question: Analyze the Text and the Caption, which named entities and their corresponding types are included in the Text? Answer: 1. ObamaCare is an other and is likely to be an ordinance 2. GOP is an organization and is likely to be a political party</p> <p>Text: # BPLStorySoFar Nemanja Matic of @ ChelseaFC has won the most tackles in the # BPL 2014/15 to date . . . Caption: the top 10 players in the premier league Question: Analyze the Text and the Caption, which named entities and their corresponding types are included in the Text? Answer:</p>
Grounded Prompt
<p>Coarse-grained: person . building . organization . location . event . art . product . Fine-grained: athlete . sports event . team . Nouns: players . premier league .</p>

LLMs, we design a heuristic example selection method, which can reduce labor costs and yield more effective examples.

Initially, we construct a set of sequences P by concatenating the text and image caption derived from the training set. Subsequently, these sequences in P are encoded with the pre-trained Sentence-BERT [21] to obtain their respective sentence embeddings. Next, we apply k-means clustering algorithm to these embeddings in order to obtain k distinct clusters, denoted as $\mathbf{p} = \{\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(k)}\}$. Each cluster has multiple sequences, $\mathbf{p}^{(i)} = [p_1^{(i)}, p_2^{(i)}, \dots]$. Within in each cluster, these sequences are sorted in ascending order based on the distance to the cluster center.

Inspired by [38], we employ heuristic criteria for sampling demonstrations. For each cluster, we iterate through the sequences and append prompt words to encourage ChatGPT to generate possible entity types in a zero-shot manner. If the generated answer contains ground truth fine-grained entity types, this example is added to the demonstration list, and the enumeration process is halted. This operation is performed for each cluster, yielding k examples. The heuristic

example selection process is shown in Algorithm 1.

After obtaining the demonstration examples, we construct the knowledgeable prompt to encourage ChatGPT to generate guiding knowledge. We design a prompt head to describe the task and important notes. The demonstration examples are filled with answers generated by zero-shot CoT method, whereas the input example remains the answer slot blank. Then we concatenate the prompt head, demonstration examples, and input example as the final prompt to query ChatGPT for guiding knowledge (see Table 2).

4.4 Grounded Prompt

In this section, we elaborate on the design of the grounded prompt to get more related candidate objects, aimed at capturing potential clues between entity and grounded object.

To ensure a comprehensive coverage of candidate objects, our grounded prompt incorporates three components. Firstly, to detect the possible objects related to the semantic information of the image, we extract nouns from image caption with NLTK toolkit. Secondly,

Algorithm 1 Heuristic Example Selection

Input: A set of paired text and image caption P and the number of example k

Output: Demonstration example list $s = [s^{(1)}, \dots, s^{(k)}]$

```

1: for  $i = 1, \dots, |P|$  do
2:    $H_i = \text{Sentence-BERT}(P_i)$ 
3: end for
4:  $p = k\text{-means}(H, k)$ 
5: for  $i = 1, \dots, k$  do
6:   Sort  $p^{(i)} = [p_1^{(i)}, p_2^{(i)}, \dots]$  in the ascending order of the distance to the cluster center
7:   for  $p_j^{(i)} \in p^{(i)}$  do
8:      $a_j^{(i)} = \text{Zero-Shot-CoT}(p_j^{(i)})$ 
9:     if  $a_j^{(i)}$  contains correct fine-grained type then
10:      Add  $s^{(i)} = [p_j^{(i)}, a_j^{(i)}]$  to  $s$ 
11:     break
12:   end if
13: end for
14: end for
15: return  $s$ 

```

to predict the possible objects related to the semantic information of the text, we extract fine-grained entity types from guiding knowledge. Thirdly, to widen the scope of candidate objects, we incorporate all the coarse-grained entity types. We concatenate these three components to form the grounded prompt (see Table 2).

Upon constructing the grounded prompt, we apply an open-set object detector Grounding DINO [16] as VGM to detect related objects from image. These objects are ranked according to their text and box probabilities, retaining the top- K as candidate objects. The corresponding features from Grounding DINO are used as object representations, and a linear projection layer is added to map the object representations to the dimensions of text embeddings.

4.5 Model Training and Prediction

The overall input to the Transformer-based model consists of four aspects: text T , guiding knowledge A , image caption C , and candidate objects V . The calculation of Transformer can be formulated as:

$$\begin{aligned} H_e &= \text{Encoder}([T; A; C; V]) \\ H_d &= \text{Decoder}(Y, H_e) \end{aligned} \quad (3)$$

where Y is the target output sequence introduced in Section 4.2. The model parameters are optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_t = - \sum_{i=1}^L \log p(Y_i | Y_{<i}, [T; A; C; V]) \quad (4)$$

where L is the length of the output sequence.

Additionally, we add a supervision measure to guide the entity visual grounding. Specifically, we calculate Intersection over Union (IoU) scores between top- K candidate objects and annotated ground truth bounding boxes. Scores falling below the threshold of 0.5 are set to 0, while the remainder are normalized. In this way, each grounded entity has a K -dimensional vector G as supervision signal.

During the training stage, if an entity is grounded in the image, we average the representations of three tokens, i.e., *in the image*, which is taken from the output of the Transformer decoder. Then we compute the similarity between this aggregated representation and

Table 3. Statistics of the Twitter dataset.

Split	#Twitter	#Entity	#Grounded Entity	#Box
Train	7000	11779	4733	5723
Dev	1500	2450	991	1171
Test	1500	2543	1046	1254

the representations of the top- K candidate object, thus deriving a visual probability distribution $p(g)$.

We adopt KL Divergence loss between $p(g)$ and G as an objective to promote entity visual grounding:

$$\mathcal{L}_v = \sum_{i=1}^N G_i \log \frac{G_i}{p(g_i)} \quad (5)$$

where N denotes the number of entities grounded in the image.

The overall training loss for our model can be formulated as:

$$\mathcal{L} = \mathcal{L}_t + \lambda * \mathcal{L}_v \quad (6)$$

In the prediction stage, we first split the output sequence with $[SEP]$ to obtain several sub-sequences. Then we extract the entity name, fine-grained entity type, and whether this entity is grounded in the image from each sub-sequence. For the entity that is grounded in the image, we further recognize the bounding box with maximal probability in $p(g)$ as the predicted result.

5 Experiments

5.1 Experimental Settings

5.1.1 Benchmark Dataset

This work used the public Twitter dataset proposed by [25] to evaluate our model. This dataset is developed from Twitter-2015 [37] and Twitter-2017 [18] by further subdividing the original entity types and adding annotations for entity visual grounding. It has 8 coarse-grained entity types and 51 fine-grained entity types. The statistics of the Twitter dataset are listed in Table 3.

5.1.2 Evaluation Metrics

We adopt precision (P), recall (R), and F1-score (F1) as evaluation metrics for this task. A predicted triple is regarded as correct only when the entity name, fine-grained entity type, and its corresponding object in the image are all correct. If the grounded entity exists, we consider the entity visual grounding result to be correct when the IoU score between the predicted bounding box and ground truth bounding box exceeds 0.5.

To prove a fair comparison with the previous method [25], we also evaluate our method on the following two subtasks: fine-grained multimodal named entity recognition (FMNER) and entity extraction with grounding (EEG). FMNER evaluates whether the model can extract both the entity name and its fine-grained entity type correctly. EEG evaluates whether the model can extract the entity name and its grounded entity in the image accurately.

5.1.3 Implementation Details

All experiments are implemented on an NVIDIA RTX 3090 GPU with PyTorch framework. We adopt T5-BASE as the Transformer model. Besides, we employ multimodal model BLIP-2 to obtain image captions, use GPT-3.5-Turbo as LLMs, and utilize Grounding

Table 4. Overall performance compared to the state-of-the-art methods on the test set. P, R, and F1 denote precision (%), recall (%), and F1-score (%). The best results are denoted in bold.

Modality	Method	FMNER			EEG			FMNERG		
		P	R	F1	P	R	F1	P	R	F1
T	HBiLSTM-CRF-Tag	62.31	56.55	59.29	49.25	43.27	46.07	34.86	32.38	33.57
	BERT-Tag	58.91	60.05	59.47	46.27	47.65	46.94	33.28	34.28	33.77
	BERT-CRF-Tag	60.06	61.38	60.72	46.93	48.44	47.67	34.41	35.51	34.95
	T5-Gen	64.83	65.32	65.07	49.03	48.91	48.97	37.38	37.29	37.33
	ChatGPT-Gen	38.04	41.99	39.92	46.28	51.09	48.56	24.10	26.61	25.30
	MVP-Text(Ours)	68.40	68.80	68.60	52.95	53.26	53.10	41.71	41.95	41.83
T+V	GVATT-EG	63.08	57.85	60.35	55.27	53.46	54.35	42.02	38.75	40.32
	UMT-EG	61.24	62.01	61.63	53.58	55.32	54.43	40.67	41.99	41.32
	UMGF-EG	61.68	61.90	61.79	54.51	55.00	54.75	41.73	42.11	41.92
	ITA-EG	63.8	62.64	63.21	57.63	56.90	57.26	43.05	42.51	42.78
	H-Index	65.25	64.45	64.84	60.82	60.10	60.46	46.83	46.28	46.55
	MMT5-EG	66.46	66.77	66.61	58.35	58.01	58.18	45.35	45.08	45.21
	TIGER	64.43	65.40	64.91	62.44	61.49	61.96	47.57	46.85	47.20
	MVP(Ours)	69.32	71.37	70.33	64.44	66.35	65.38	52.27	53.82	53.03

Table 5. Performance on FMNERG according to coarse-grained entity type. We choose the three most and three least types for display.

Method	Per.	Loc.	Build.	Org.	Prod.	Other
GVATT-EG	35.21	61.64	35.37	42.60	15.38	41.03
UMT-EG	37.10	63.58	35.09	42.82	18.28	38.24
UMGF-EG	37.04	63.16	38.51	44.71	17.39	38.89
ITA-EG	37.91	65.52	39.16	44.34	17.18	36.36
H-Index	45.13	62.33	32.88	46.68	28.19	41.81
MMT5-EG	38.61	69.44	37.18	46.30	16.18	46.98
TIGER	43.78	67.69	40.00	46.75	27.38	48.28
MVP(Ours)	53.91	68.85	40.26	49.08	31.95	47.37

Table 6. Experimental results of ablation study on the test set.

Method	FMNER	EEG	FMNERG
MVP	70.33	65.38	53.03
-w/o Guiding	68.53	63.89	49.26
-w/o Caption	71.13	61.81	49.28
-w/o Object	68.60	53.10	41.83
-w/o ICL	68.92	62.55	50.35
-w/o CoT	67.68	65.19	51.48
-w/o Noun	68.05	64.25	51.42
-w/o Fine	69.31	64.58	52.13
-w/o Coarse	69.02	63.18	50.85

Table 7. Effect of whether entity is grounded in the image.

Ground	Method	FMNER	EEG	FMNERG
False	TIGER	59.24	69.58	51.17
	MVP	63.06	76.01	56.16
True	TIGER	68.01	57.19	45.22
	MVP	74.24	61.05	51.35

DINO to detect candidate objects. We train our model for 10 epochs with a batch size of 16 and a learning rate of 1e-4. The number of demonstration examples k is set to 8. The maximum number of candidate objects K is set to 30 and λ is set to 1.0. We adopt AdamW optimizer to minimize the loss function.

5.1.4 Baseline Model and Variants

The compared methods can be divided into two categories according to the modality used: unimodal methods (i.e., only use text modality and set grounded entity to *None*) and multimodal methods.

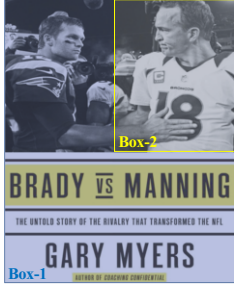

(1) Unimodal methods: **HBiLSTM-CRF-Tag** is a sequence labeling method with a hierarchical BiLSTM-CRF architecture [18]. **BERT-Tag** and **BERT-CRF-Tag** are variant models of HBiLSTM-CRF-Tag with BERT and BERT-CRF architectures, respectively. **T5-Gen** employs a generative sequence-to-sequence model to extract entity names and fine-grained entity types. **ChatGPT-Gen** means directly using ChatGPT to generate the answers with knowledgeable prompt.

(2) Multimodal methods: **GVATT-EG** is a sequence labeling BiLSTM-CRF method [18] that integrates visual features and an entity grounding model is added to achieve entity grounding in the image. **UMT-EG**, **UMGF-EG**, and **ITA-EG** use the MNER methods from [33], [35], and [28] respectively. And an entity grounding model is also added to them. **MMT5-EG** is the combination of T5-Tag and entity grounding model. **H-index** [34] and **TIGER** [25] adopt an end-to-end generation framework to achieve FMNERG task with BART and T5 architecture, respectively.

5.2 Main Results

The main results are reported in Table 4 and the performance (F1-score) according to coarse-grained entity type is presented in Table 5. We can observe that: (1) MVP achieves the best performance on the FMNERG task and its two subtasks, outperforming previous state-of-the-art TIGER by a large margin. This significant improvement demonstrates the effectiveness of the proposed method. (2) The multimodal methods generally deliver superior performance compared to unimodal methods. This suggests the visual modality and entity grounding model can benefit the entity extraction with grounding task, thus boosting model performance. (3) The generation-based methods (e.g., T5-GEN, MMT5-EG) usually drive better performance than the corresponding tagging-based methods. The reason may be that these generation-based methods excel at the FMNER task. (4) The end-to-end methods (e.g., TIGER, MVP) hold a clear advantage over the pipeline approaches. We attribute the reason to the fact that the pipeline methods tend to suffer from error propagation, thus performing worse on the EEG task. In contrast, our model

Table 8. Case study on the test set. Red denotes wrong predictions.

My Brady vs. Manning book will be released on Sept. 22. Pre-order on Amazon ht...		Seth Meyers has a message for Hillary voters: Bernie is not hurting your candidate.	
	<p><i>Ground Truth</i> (Brady vs. Manning, written work, Box-1) (Amazon, website, None)</p> <p><i>TIGER</i> (Brady, athlete, Box-2) ✗ (Manning, athlete, Box-2) ✗ (Amazon, website, None) ✓</p> <p><i>MVP</i> (Brady vs. Manning, written work, Box-1) ✓ (Amazon, website, None) ✓</p>		<p><i>Ground Truth</i> (Seth Meyers, actor, Box-1) (Hillary, politician, None) (Bernie, politician, Box-2)</p> <p><i>TIGER</i> (Seth Meyers, actor, Box-1) ✓ (Hillary, politician, Box-1) ✗ (Bernie, politician, None) ✗</p> <p><i>MVP</i> (Seth Meyers, actor, Box-1) ✓ (Hillary, politician, None) ✓ (Bernie, politician, Box-2) ✓</p>

can optimize the two subtasks in an end-to-end manner and combine more effective knowledge and more relevant candidate objects, leading to better performance. (5) Among methods that rely solely on text modality, MVP-TEXT achieves the best performance and even outperforms some multimodal methods (e.g., GVATT-EG), especially on FMNER. This indicates the proposed method exhibits the robust ability to extract entity names and fine-grained entity types. (6) MVP also outperforms previous methods on the majority of coarse-grained entity types. It exceeds the previous SOTA method TIGER by more than 10% on *person* type. This suggests that the guiding knowledge is more effective for *person* than other types.

5.3 More Analysis

Effect of Transformer Input. To assess the effect of each input to the Transformer, we undertake ablation studies by comparing MVP with a series of variant methods. The results are reported in Table 6. We can observe that: (1) After removing guiding knowledge, the model performance drops significantly, which suggests guiding knowledge plays an important role in FMNER and EEG tasks. (2) The elimination of image captions from the input sequence results in a considerable performance drop of the EEG, illustrating that the caption can promote the understanding of the image and aid in grounding entity to object in the image. (3) The exclusion of candidate objects leads to a decrease of 12.28% in the F1-score on the EEG task. This demonstrates the vital contribution of visual modality in EEG.

Effect of Knowledgeable Prompt. To investigate the effect of each component in constructing knowledgeable prompt, we compare MVP with two variant knowledgeable prompts. The results are shown in Table 6. *-w/o ICL* and *-w/o CoT* represent removing demonstration examples and CoT trigger words from knowledgeable prompt, respectively. We can find that using zero-shot CoT prompt achieves poor performance. This is because the demonstration examples can help LLMs understand and execute the task more effectively, thus improving model performance. Without CoT trigger words, the model performance also drops. This is because the CoT trigger words can encourage the model to consider the entity type from coarse-grained to fine-grained level, resulting in more accurate guiding knowledge.

Effect of Grounded Prompt. To evaluate the effect of each component in constructing grounded prompt, we compare MVP with three variant grounded prompts. The results are reported in Table 6. *-w/o Noun*, *-w/o Fine*, and *-w/o Coarse* denote removing nouns, fine-grained entity type, and coarse-grained entity type from grounded

prompt, respectively. We can observe that removing each component leads to performance decay, which indicates each component contributes significantly to grounded prompt. After removing coarse-grained entity types, the model performance drops the most. The reason may be that the visual grounding model can understand coarse-grained entity types better and provide more accurate object detection results.

Effect of Whether Grounded in Image. In this section, we study the effect of the presence or absence of grounded entities in the image. The test set is divided into two subsets: *False* denotes no grounded entities and *True* means existing grounded entities. The results are reported in Table 7. We find that MVP performs generally better than TIGER on two subsets, which demonstrates the superiority of MVP. Besides, the methods on *True* subset achieve better performance on FMNER but poorer performance on EEG. We attribute the reason to the fact that the training process of entity grounding can promote the FMNER task.

Case Study. In this section, we conduct case studies to further illustrate the effectiveness of MVP and show the results in Table 8. For the first case, TIGER fails to predict entity *Brady vs. Manning* as a written work. It thinks *Brady* and *Manning* are athletes and grounded them to objects in the image wrongly. This suggests TIGER lacks knowledge of *Brady vs. Manning* and may not detect the book object, leading to wrong predictions. For the second case, TIGER confuses the objects corresponding to *Hillary* and *Bernie*. The reason may be the shortcomings of correlations between the entity and its corresponding object. However, MVP can equip the model with rich knowledge and mine the implicit clues between entities and objects.

6 Conclusion

In this paper, we present a Multi-view Prompt (MVP) method for the FMNER task, which employs knowledgeable prompt and grounded prompt to collaborate with LLMs and VGMs. The knowledgeable prompt leverages a chain-of-thought method to guide LLMs to think entity type from coarse-grained to fine-grained level and adopt a heuristic approach to select demonstration examples for in-context learning. The grounded prompt can incorporate information from image caption and guiding knowledge to detect potential objects, which can enhance correlations between entities and objects. Experimental results illustrate that MVP achieves the best performance among a series of baseline models.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [4] L. Chen, W. Ma, J. Xiao, H. Zhang, and S.-F. Chang. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1036–1044, 2021.
- [5] S. Chen, G. Aguilar, L. Neves, and T. Solorio. Can images help recognize entities? A study of the role of images for multimodal NER. In W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, editors, *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT 2021, Online, November 11, 2021*, pages 87–96. Association for Computational Linguistics, 2021.
- [6] X. Chen, N. Zhang, L. Li, Y. Yao, S. Deng, C. Tan, F. Huang, L. Si, and H. Chen. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv preprint arXiv:2205.03521*, 2022.
- [7] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.
- [8] M. Jia, L. Shen, X. Shen, L. Liao, M. Chen, X. He, Z. Chen, and J. Li. Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8032–8040, 2023.
- [9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [10] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [11] J. Li, H. Li, Z. Pan, D. Sun, J. Wang, W. Zhang, and G. Pan. 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, 2023.
- [12] J. Liu, Z. Zhang, Z. Guo, L. Jin, X. Li, K. Wei, and X. Sun. Emotion-cause pair extraction with bidirectional multi-label sequence tagging. *Appl. Intell.*, 53(24):30400–30415, 2023. doi: 10.1007/S10489-023-05140-Z. URL <https://doi.org/10.1007/s10489-023-05140-z>.
- [13] J. Liu, Z. Zhang, Z. Guo, L. Jin, X. Li, K. Wei, and X. Sun. KEPT: knowledge enhanced prompt tuning for event causality identification. *Knowl. Based Syst.*, 259:110064, 2023. doi: 10.1016/J.KNOSYS.2022.110064. URL <https://doi.org/10.1016/j.knosys.2022.110064>.
- [14] J. Liu, Z. Zhang, K. Wei, Z. Guo, X. Sun, L. Jin, and X. Li. Event causality extraction via implicit cause-effect interactions. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6792–6804. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.420. URL <https://doi.org/10.18653/v1/2023.emnlp-main.420>.
- [15] J. Liu, K. Wei, and C. Liu. Multimodal event causality reasoning with scene graph enhanced interaction network. In M. J. Wooldridge, J. G. Dy, and S. Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 8778–8786. AAAI Press, 2024. doi: 10.1609/AAAI.V38I8.28724. URL <https://doi.org/10.1609/aaai.v38i8.28724>.
- [16] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [17] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, and D. S. Rosenblum. Mmkg: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 459–474. Springer, 2019.
- [18] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, 2018.
- [19] S. Moon, L. Neves, and V. Carvalho. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, 2018.
- [20] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [21] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [22] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [23] O. Rubin, J. Herzig, and J. Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021.
- [24] Z. Shao, Z. Yu, M. Wang, and J. Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983, 2023.
- [25] J. Wang, Z. Li, J. Yu, L. Yang, and R. Xia. 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, 2023.
- [26] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*, 2021.
- [27] X. Wang, J. Cai, Y. Jiang, P. Xie, K. Tu, and W. Lu. Named entity and relation extraction with multi-modal retrieval. *arXiv preprint arXiv:2212.01612*, 2022.
- [28] X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, and K. Tu. ITA: image-text alignments for multi-modal named entity recognition. In M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3176–3189. Association for Computational Linguistics, 2022.
- [29] X. Wang, W. Zhu, and W. Y. Wang. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, 2023.
- [30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [31] S. Yang, G. Li, and Y. Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961, 2020.
- [32] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019.
- [33] J. Yu, J. Jiang, L. Yang, and R. Xia. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics, 2020.
- [34] J. Yu, Z. Li, J. Wang, and R. Xia. 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, 2023.
- [35] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou. 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, 2021.
- [36] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.
- [37] Q. Zhang, J. Fu, X. Liu, and X. Huang. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [38] Z. Zhang, A. Zhang, M. Li, and A. Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.
- [39] F. Zhao, C. Li, Z. Wu, S. Xing, and X. Dai. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal ner. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3983–3992, 2022.
- [40] C. Zheng, Z. Wu, T. Wang, Y. Cai, and Q. Li. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 23:2520–2532, 2020.