MAKAR: a Multi-Agent framework based Knowledge-Augmented Reasoning for Grounded Multimodal Named Entity Recognition

Xinkui Lin[1,2,3]*, Yuhui Zhang[1,2], Yongxiu Xu[1,2]†, Kun Huang[3], Hongzhang Mu[1,2], Yubin Wang[1,2], Gaopeng Gou[1,2], Li Qian [3], Li Peng[3], Wei Liu[3], Jian Luan[3], Hongbo Xu[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]MiLM Plus, Xiaomi Inc, China
{linxinkui,xuyongxiu}@iie.ac.cn

Abstract

Grounded Multimodal Named Entity Recognition (GMNER), which aims to extract textual entities, their types, and corresponding visual regions from image-text data, has become a critical task in multimodal information extraction. However, existing methods face two major challenges. First, they fail to address the semantic ambiguity caused by polysemy and the long-tail distribution of datasets. Second, unlike visual grounding which provides descriptive phrases, entity grounding only offers brief entity names which carry less semantic information. Current methods lack sufficient semantic interaction between text and image, hindering accurate entity-visual region matching. To tackle these issues, we propose MAKAR, a Multi-Agent framework based Knowledge-Augmented Reasoning, comprising three agents: Knowledge Enhancement, Entity Correction, and Entity Reasoning Grounding. Specifically, in the named entity recognition phase, the Knowledge Enhancement Agent leverages a Multimodal Large Language Model (MLLM) as an implicit knowledge base to enhance ambiguous image-text content with its internal knowledge. For samples with low-confidence entity boundaries and types, the Entity Correction Agent uses web search tools to retrieve and summarize relevant web content, thereby correcting entities using both internal and external knowledge. In the entity grounding phase, the Entity Reasoning Grounding Agent utilizes multi-step Chain-of-Thought reasoning to perform grounding for each entity. Extensive experiments show that MAKAR achieves state-of-the-art performance on two benchmark datasets. Code is available at: https://github.com/Nikol-coder/MAKAR.

1 Introduction

Multimodal Named Entity Recognition (MNER) (Chen et al., 2022; Chen and Feng, 2023; Liu et al., 2022; Yuan et al., 2023), a pivotal task in natural language processing, has propelled the advancement of knowledge graphs. With the development of multimodal knowledge graphs, MNER has expanded into Grounded Multimodal Named Entity Recognition (GMNER) (Yu et al., 2023), which identifies named entities from image-text pairs, along with their categories and corresponding visual region coordinates. Furthermore, Fine-grained Multimodal Named Entity Recognition and Grounding (FMNERG) (Wang et al., 2023a) refines textual entity categorization into more detailed classes.

Figure 1: Illustrations of MNER, GMNER, and FMNERG tasks.

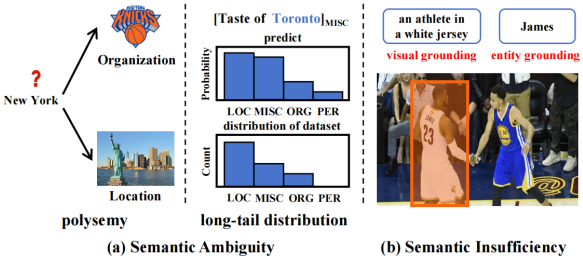

Figure 2: Key issues.

Recent methods (Li et al., 2024a; Wang et al., 2024; Tang et al., 2025) have made progress in this field, but have overlooked two key issues (shown in Fig. 2). One is semantic ambiguity arising from polysemy and the long-tail distribution in datasets. The other is that, unlike visual grounding which

* Work done during an internship at Xiaomi Inc.
† Corresponding Author: Yongxiu Xu.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6122–6142
November 4-9, 2025

provides a full phrase, entity grounding often offers brief entity name, which contain less semantic information. Existing entity grounding methods (Girshick, 2015; Zhang et al., 2021b) lack sufficient semantic interaction between text and image, hindering the accurate identification of the corresponding visual regions.

To tackle these challenges, we propose a Multi-Agent framework based Knowledge-Augmented Reasoning, termed MAKAR. It features three agents: Knowledge Enhancement Agent, Entity Correction Agent, and Entity Reasoning Grounding Agent, which interact over multiple rounds to perform precise grounded multimodal named entity recognition. Knowledge Enhancement Agent (KEA) leverages its inherent knowledge to enhance each input sample, generating preliminary candidate entities. Entity Correction Agent (ECA) utilizes external knowledge to revise samples with low-confidence in entity boundaries and types. By invoking web search to retrieve and summarize relevant background knowledge, it refines these low-confidence samples to obtain final textual entities. Entity Reasoning Grounding Agent (ERGA) conducts progressive reasoning for each textual entity, fully exploiting the augmented knowledge and image-text information to execute entity grounding. Our MAKAR framework effectively improves both the accuracy and explainability of GMNER.

The main contributions of our work can be summarized as follows:

- We propose a multi-agent framework based knowledge-augmented reasoning that integrates the internal knowledge of MLLMs with external web-retrieved knowledge, resolving polysemy and long-tail distribution challenges in named entity recognition.
- We design a progressive reasoning grounding approach combining SFT-initialization with GRPO-based policy optimization, achieving alignment between entities and visual regions.
- Extensive experiments on GMNER and FMNERG datasets demonstrate that our framework achieves state-of-the-art performance.

2 Related work

Grounded Multimodal Named Entity Recognition. Grounded Multimodal Named Entity Recognition (GMNER) extracts textual entities, their types, and corresponding visual region coordinates from image-text pairs. Early methods like H-index (Yu et al., 2023) and Tiger (Wang et al., 2023a) used existing object detectors (Girshick, 2015) to detect visual objects in advance. Then, by treating both text and visual objects as inputs, they generated triplets like *(textual entity, entity type, visual region)* through a generative framework. With the development of Multimodal Large Language Models (MLLMs), GEM (Wang et al., 2024) and RiVEG (Li et al., 2024a) leverage MLLMs to improve textual knowledge and grounding. However, these approaches overlook two critical issues: (1) semantic ambiguity caused by polysemy and long-tail distribution in datasets; and (2) insufficient entity-visual region interaction due to the limited semantic information of entities.

Our work addresses these issues by using MLLMs to retrieve internal and external knowledge, resolve semantic ambiguity, and correct prediction biases. Meanwhile, we enrich the semantic information of entities through leveraging existing knowledge and adopt progressive reasoning grounding to improve the interpretability and accuracy of entity grounding.

3 Method

In this section, we first introduce the GMNER task, and then explain our MAKAR framework (shown in Fig. 3) in detail. Our framework consists of three agents: (1) *Knowledge Enhancement Agent (KEA, §3.2)* references similar samples and provides auxiliary information for input data from its own knowledge base to perform knowledge enhancement, yielding preliminary candidate entities; (2) *Entity Correction Agent (ECA, §3.3)* is responsible for conducting external knowledge retrieval on low-confidence candidate entities, expanding background knowledge, and correcting the types and boundaries of entities; (3) *Entity Reasoning Grounding Agent (ERGA, §3.4)* integrates existing knowledge to enrich the semantic information of textual entities, adopting progressive reasoning grounding to enhance the interpretability of entity grounding.

3.1 Task Definitions

Given a sentence T and its accompanying image I, the GMNER task aims to extract and classify textual entities in T, then locate their coordinates in I. Outputs can be represented as a set of triples:

$$F = \{(e_1, t_1, c_1), ..., (e_N, t_N, c_N)\} \quad (1)$$

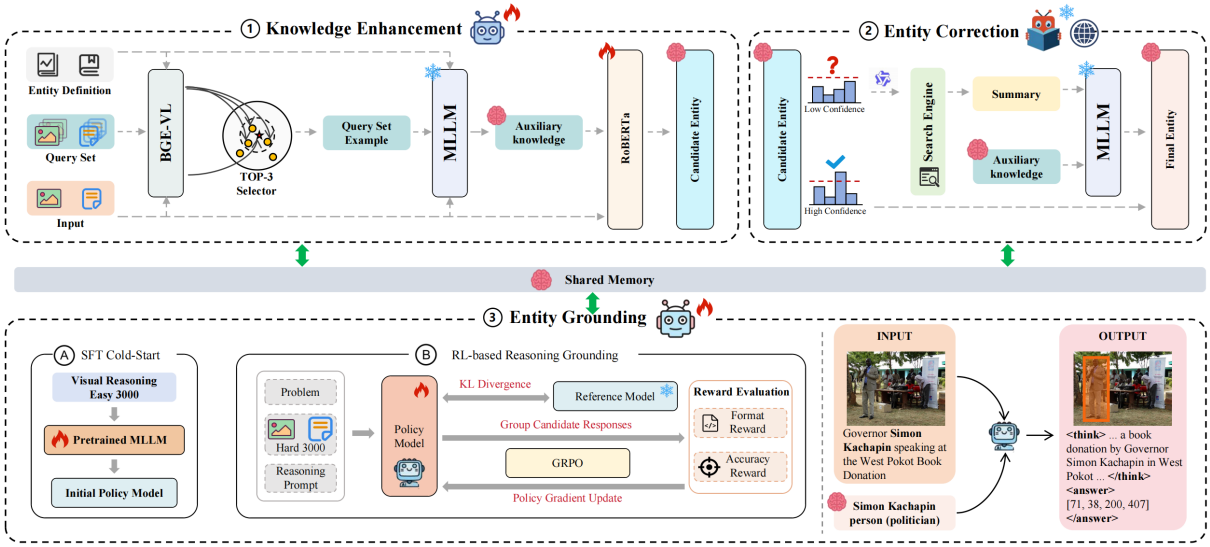

Figure 3: The overall framework of our MAKAR.

where N represents the total number of textual entities in T, e_i denotes the i-th textual entity in T, t_i denotes the type of e_i, and c_i denotes the visual region corresponding to textual entity e_i. If e_i has a corresponding visual region in I, c_i is a four-dimensional vector containing the coordinates of the bounding box; otherwise, c_i is None, denoted as (0,0,0,0). c_i can be expressed as:

$$c_i = \begin{cases} (x_1^i, y_1^i, x_2^i, y_2^i), & \text{if } e_i \text{ is grounded} \\ (0,0,0,0), & \text{if } e_i \text{ is ungrounded} \end{cases} \quad (2)$$

where (x_1^i, y_1^i) and (x_2^i, y_2^i) denote the top-left and bottom-right coordinates in I.

3.2 Knowledge Enhancement Agent

The Knowledge Enhancement Agent (KEA) provides auxiliary knowledge for preliminary named entity recognition and feeds the results back to the Entity Correction Agent. First, we design distinct prompt templates for coarse-grained and fine-grained entity types, explicitly defining each entity category to clarify their semantic boundaries. Then, to better obtain auxiliary knowledge, we cluster training samples using K-Means (Nie et al., 2023) based on entity type and multimodal representations, selecting 200 representative samples as the Query Set for knowledge enhancement. Next, we use a Multimodal Large Language Model (MLLM) to enrich the knowledge of entities in the Query Set. After that, we compute cross-modal similarity between the input sample and the Query Set using BGE-VL (Zhou et al., 2024), picking the top-3 most similar query set samples as prompt examples. The Knowledge Enhancement Agent leverages these examples, entity type definitions, and input samples to generate auxiliary knowledge for each input. The auxiliary knowledge is concatenated with the original text using a separator token, forming the augmented input X, which is formalized as:

$$X = (s_1, \ldots, s_{N_1}, \langle \text{SEP} \rangle, aux_1, \ldots, aux_{N_2}) \quad (3)$$

where s_i and aux_i denote tokens from the input text and auxiliary knowledge, respectively. This augmented input X is encoded by a fine-tuned RoBERTa (Li et al., 2023, 2024a) model, whose embeddings are fed into a Conditional Random Field (CRF) (Huang et al., 2015) layer to decode the predicted label sequence y. The probability of y given X is calculated as:

$$P(y|X) = \frac{\prod_{i=1}^{N_1+N_2+1} \exp(\psi(y_{i-1}, y_i, x_i))}{\sum_{y' \in Y} \prod_{i=1}^{N_1+N_2+1} \exp(\psi(y'_{i-1}, y'_i, x_i))} \quad (4)$$

where Y is the set of all possible label sequences given the input X. The potential function ψ decomposes into transition and emission components:

$$\psi(y_{i-1}, y_i, \mathbf{x}_i) = \underbrace{T_{y_{i-1}, y_i}}_{\text{transition potential}} + \underbrace{\mathbf{W}_s \mathbf{h}_i + b_s}_{\text{emission potential}} \quad (5)$$

The transition potential T_{y_{i-1}, y_i} denotes a learnable transition matrix, and the emission potential $\boldsymbol{W}_s \boldsymbol{h}_i + b_s$ represents the label scores derived from RoBERTa embeddings $\boldsymbol{h}_i \in \mathbb{R}^d$ via linear projection. Here, $\boldsymbol{W}_s \in \mathbb{R}^{K \times d}$ is a learnable weight

matrix,$b_s \in \mathbb{R}^K$ is the bias term, and K is the number of label classes. Ultimately, we utilize the negative log-likelihood loss function to train the model:

$$\mathcal{L}_{\text{MNER}} = -\log P(y^*|X) \tag{6}$$

where y^* denotes ground-truth entity type labels for the text sequences. KEA effectively addresses semantic ambiguity caused by polysemy through the internal knowledge of MLLM.

3.3 Entity Correction Agent

Wrong example	True example
the [black eyed peas]$_{\text{ORG}}$	[the black eyed peas]$_{\text{ORG}}$
Taste of [Toronto]$_{\text{LOC}}$	[Taste of Toronto]$_{\text{MISC}}$
[Switzerland Co]$_{\text{ORG}}$	[Switzerland Co]$_{\text{LOC}}$

Table 1: Examples of type and boundary error. The incorrect annotations are in red [*], while the correct annotations are marked in blue [*].

The Entity Correction Agent (ECA) refines low-confidence entities by integrating external knowledge with results from KEA. Owing to long-tail distributions in datasets, certain candidate entities exhibit boundary or type errors, resulting in low-confidence scores. ECA dynamically decides its next action based on the confidence scores delivered by KEA (Yao et al., 2023). Specifically, for each candidate entity we compute boundary probability p_b and type probability p_t and compare them against a predefined threshold τ. For low-confidence entities ($\min(p_b, p_t) < \tau$), the ECA uses web search tools to retrieve 3 most relevant web pages based on the input text, then summarizes them into distilled background knowledge $Back$. Meanwhile, leveraging this background knowledge $Back$, results Res from KEA, the input image I, and text T, the ECA refines entity boundaries and types. For high-confidence entities, ECA takes them as final. This process is formalized as:

$$E^{\text{Final}} = \begin{cases} \text{MLLM}_{\text{Correct}}(Res, Back, I, T) & \text{if } \min(p_b, p_t) < \tau, \\ Res & \text{if otherwise} \end{cases} \tag{7}$$

where E^{Final} denotes the corrected entity set. By dynamically leveraging external knowledge to correct entity boundaries and types, ECA effectively addresses semantic ambiguity induced by long-tail distributions.

3.4 Entity Reasoning Grounding Agent

The Entity Reasoning Grounding Agent (ERGA) combines internal knowledge from the KEA and external knowledge from ECA to infer the visual coordinates corresponding to each entity in the image. Unlike traditional visual grounding tasks that provide full descriptive phrases, entity grounding only provides brief entity names, which contain less semantic information. This restriction limits the effectiveness of traditional visual grounding methods (Li et al., 2024b).

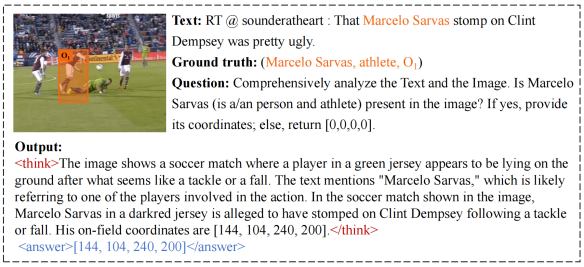

Figure 4: Example of Entity Reasoning Grounding.

To tackle this challenge, we propose Entity Reasoning Grounding (Fig. 4). It enriches entity names by integrating textual and visual cues, generating detailed descriptions with step-by-step reasoning during grounding. Inspired by DeepSeek R1 (DeepSeek-AI et al., 2025), we train the Entity Reasoning Grounding Agent (ERGA) in two phases (as shown in Fig. 3): (1) SFT Cold-Start: Initialize the reference model via Supervised Fine-Tuning (SFT) using multi-step Chain-of-Thought (CoT) data from simple samples. (2) RL-based Reasoning Grounding: Perform Group Relative Policy Optimization (GRPO)-based reinforcement fine-tuning on hard samples to further improve the model's reasoning grounding performance.

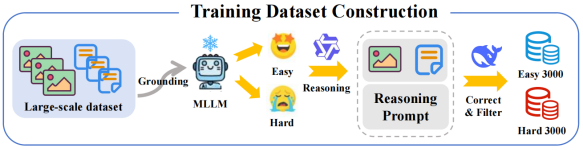

Figure 5: Dataset Construction.

To enable ERGA to learn distinct reasoning capabilities across training stages, we categorize the training data into simple and hard subsets based on grounding difficulty (Cheng et al., 2025). We first prompt Qwen2.5-VL-7B (Bai et al., 2025) to directly generate entity coordinates for image-text pairs in the training set. Samples are classified as easy if the predicted bounding boxes achieve an Intersection over Union (IoU) > 0.5 with the ground truth; otherwise, they are labeled hard. Next, we use Qwen2.5-VL-32B to generate CoT

reasoning processes for both subsets. These CoT outputs are further refined and filtered by reasoning-capable LLMs (QwQ (Team, 2025) and DeepSeek R1) to remove incorrect or low-quality inferences, ensuring logical consistency and semantic accuracy. Through this automated pipeline, we construct 3,000 easy CoT samples for SFT cold start and 3,000 hard CoT samples for RL-based reasoning.

3.4.1 SFT Cold-Start

During the cold start phase, we conduct supervised fine-tuning of the MLLM on easy samples for entity reasoning grounding, endowing it with the ability to generate multi-step CoT reasoning. Specifically, MLLM first extracts contextual knowledge about the target entity from the image-text pair, then identifies the corresponding visual region, and finally predicts the bounding box coordinates. Each sample is formalized as (I, T, Q, O), where I is the input image, T denotes the input text, Q denotes the question including the target entity, and O aggregates both the reasoning and the final prediction coordinates. The training objective is to maximize the likelihood of generating O with coherent reasoning and accurate coordinates given (I, T, Q), effectively transforming concise entity names into detailed visual descriptions through multi-step CoT reasoning. The SFT loss function is defined as:

$$\mathcal{L}_{SFT} = \mathbb{E}\left[\log \pi_\theta(O|I, T, Q)\right] \tag{8}$$

where θ denotes the parameters of the MLLM, and $\pi_\theta(O \mid I, T, Q)$ represents the conditional probability of generating the output O. The resulting model π_{SFT} initializes the next stage, setting the stage for subsequent reinforcement learning.

3.4.2 RL-based Reasoning Grounding

Reward Evaluation. For entity reasoning grounding, we design two types of rewards: Format Reward r_{format} and Accuracy Reward r_{acc}. Format Reward ensures structured CoT responses with labels ⟨think⟩ for reasoning and ⟨answer⟩ for grounding. If both are present and correctly formatted, $r_{\text{format}} = 1$; otherwise, $r_{\text{format}} = 0$. Accuracy Reward assesses the correctness of the answer within ⟨answer⟩ by checking if the IOU between the predicted bounding box coordinates and the ground truth exceeds a threshold of 0.5.

$$r_{\text{acc}} = \begin{cases} 1, & \text{if IOU} > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

The overall reward function r is defined as:

$$r = \alpha \cdot r_{\text{format}} + (1 - \alpha) \cdot r_{\text{acc}} \tag{10}$$

where α is a hyperparameter that balances the emphasis on structured reasoning and factual accuracy.

Policy Update with Relative Advantage. During the SFT cold start phase, ERGA learns basic reasoning patterns from easy samples. To enhance its reasoning ability on hard samples, we introduce a GRPO-based reinforcement fine-tuning method. Specifically, the policy model π_θ initialized from π_{SFT}, generates multiple candidate responses o_i for a given input.

$$o_i \sim \pi_\theta(o \mid I, T, Q), \quad \text{for} \quad i = 1, \ldots, G \tag{11}$$

Candidate responses are evaluated using predefined reward functions to obtain a reward sequence $\{r_1, \ldots, r_G\}$. These rewards are normalized to calculate relative advantages $A_1, \ldots, A_G$, defined as:

$$A_i = \frac{r_i - \text{mean}\{r_1, \ldots, r_G\}}{\text{std}\{r_1, \ldots, r_G\}} \tag{12}$$

GRPO estimates the policy update magnitude by calculating the probability ratio of each response under the new policy π_{new} relative to the old policy π_{old}:

$$\text{ratio}(i) = \frac{\pi_{\theta_{\text{new}}}(o_i|I, T, Q)}{\pi_{\theta_{\text{old}}}(o_i|I, T, Q)} \tag{13}$$

To stabilize training and avoid excessive updates, this ratio is clipped within $[1 - \epsilon, 1 + \epsilon]$. Furthermore, to explicitly regularize distributional shifts and mitigate catastrophic forgetting during reinforcement learning, we introduce a Kullback-Leibler (KL) divergence (Wu et al., 2025) penalty between the policy $\pi_{\theta_{\text{new}}}$ and the reference model π_{SFT}.

$$D_{KL}(\pi_{\theta_{new}} \| \pi_{\text{SFT}}) = \frac{\pi_{\text{SFT}}}{\pi_{\theta_{new}}} - \log\left(\frac{\pi_{\text{SFT}}}{\pi_{\theta_{new}}}\right) - 1 \tag{14}$$

The GRPO loss function is defined as:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}} = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(o|I,T,Q)} \\ \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min \Big(\text{radio(i)} \cdot A_i, \\ \text{clip}\left(\text{radio(i)}, 1 - \varepsilon, 1 + \varepsilon\right) \cdot A_i \Big) \\ - \beta \cdot D_{KL}(\pi_{\theta_{new}} \| \pi_{\text{SFT}}) \Bigg] \end{aligned} \tag{15}$$

where hyperparameter β ensures alignment with pre-trained knowledge while allowing adaptive updates. By integrating these mechanisms, GRPO achieves stable and effective policy optimization, driving our model to generate higher-quality, verifiable CoT reasoning paths.

4 Experiment

4.1 Datasets

	Twitter-GMNER			Twitter-FMNERG		
	Train	Dev	Test	Train	Dev	Test
Entity type	4	4	4	51	51	51
Tweet	7000	1500	1500	7000	1500	1500
Entity	11,782	2,453	2,543	11,779	2,450	2,543
Grounded	4,694	986	1,036	4,733	991	1,046
Box	5,680	1,166	1,244	5,723	1,171	1,254

Table 2: The statistics of two GMNER datasets.

In our work, we conduct experiments on two datasets: Twitter-GMNER and Twitter-FMNERG. Twitter-GMNER focuses on extracting four coarse-grained types of textual entities from text-image pairs. Twitter-FMNERG, built upon GMNER, expands to 8 coarse-grained and 51 fine-grained entity types. Table 2 presents the statistical details of the datasets, with more information in Appendix F.

4.2 Evaluation Metrics

Following prior work (Yu et al., 2023; Wang et al., 2023a), we evaluate our framework across three tasks: (1) Multimodal Named Entity Recognition (MNER) predicts textual entity boundaries and types. (2) Entity Extraction & Grounding (EEG) identifies textual entity boundaries and their corresponding visual regions. We use an IoU threshold of 0.5 to assess the accuracy of visual region predictions.(3) Multimodal Named Entity Recognition and Grounding (MNERG) evaluates both MNER and EEG comprehensively, ensuring triplet (e_i, t_i, c_i) accuracy. For all tasks, we use F1-score as the evaluation metric. The calculations are based on the following criteria:

For entity boundary and type correctness (C_e/C_t):

$$C_e/C_t = \begin{cases} 1, & \text{if } p_e/p_t = g_e/g_t; \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

For visual region correctness (C_c):

$$C_c = \begin{cases} 1, & \text{if } p_c = g_c = \text{None}; \\ 1, & \text{if IoU}(p_c, g_c) > 0.5; \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

$$\text{correct} = \begin{cases} 1, & \text{if } C_e \& C_t \& C_c = 1; \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

A prediction is deemed correct if and only if $C_e \& C_t \& C_c = 1$. Precision (Pre), Recall (Rec), and F1-score are calculated as follows:

$$\text{Pre} = \frac{\#\text{correct}}{\#\text{predict}}, \quad \text{Rec} = \frac{\#\text{correct}}{\#\text{gold}} \quad (19)$$

$$\text{F1} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \quad (20)$$

where #correct, #predict, and #gold denote the number of correct predictions, total predictions, and gold labels.

4.3 Implementation Details

In KEA, we employ BGE-VL-base as the multimodal encoder to compute similarities between different samples. We select different variants of Qwen and LLaMA (Grattafiori et al., 2024) as our knowledge base for knowledge enhancement. We also fine-tune RoBERTa (Zhuang et al., 2021) using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 5.0e-5 for 20 epochs. In ECA, we use Qwen2.5-VL-7B to perform web searches and summarize background knowledge for correction, with a correction threshold τ of 0.5. The MLLM used for correction is consistent with that used in KEA. In ERGA, we train the model for 5 epochs with a learning rate of 1.0e-5 during the SFT cold start phase, and for 10 epochs with a learning rate of 1.0e-6 during the RL phase. The weights of the reward function α and the KL-divergence penalty β are set to 0.5 and 0.01 respectively. Our implementation is built on the open source frameworks Easy-R1 (Sheng et al., 2024; Zheng et al., 2025) and LlamaFactory (Zheng et al., 2024), ensuring reproducibility and scalability. To ensure fair comparisons, we use the same textual encoder as other baselines. All model components run on 8×A800 GPUs using PyTorch. We conduct 3 averages for each experiment.

4.4 Baselines

To evaluate our framework, we compare against three categories of baselines: (1) Text-only models: Solely perform textual entity extraction and set all visual region predictions to (0,0,0,0). These include sequence labeling methods like BiLSTM-CRF (Lu

Modality	Methods	GMNER			FMNERG		
		MNERG	MNER	EEG	MNERG	MNER	EEG
Text	BiLSTM-CRF-None	42.07	75.58	47.49	33.57	59.29	46.07
	Bert-None	42.96	77.30	47.63	33.77	59.47	46.94
	Bert-CRF-None	43.78	77.93	48.07	34.95	60.72	47.67
	BARTNER-None	44.82	79.83	48.99	37.33	65.07	48.97
Text+Image	UMT-RCNN-EVG	50.29	78.58	54.78	41.32	61.63	45.43
	UMGF-VinVL-EVG	51.67	78.83	55.74	41.92	61.79	54.75
	ITA-VinVL-EVG	51.56	79.37	55.69	42.78	63.21	57.26
	BARTMNER-VinVL-EVG	52.45	80.39	55.66	45.21	66.61	58.18
	H-Index	56.41	79.73	61.18	-	-	-
	TIGER	-	-	-	46.55	64.91	61.96
	RiVEG	67.06	84.51	68.79	-	-	-
	GEM	61.54	84.81	64.49	52.48	70.80	65.20
	MQSPN	58.76	80.43	62.40	48.57	67.09	62.54
	SCANNER	68.52	-	-	-	-	-
	MAKAR (Ours)	**71.88**	**86.38**	**74.64**	**60.54**	**71.24**	**75.66**
	Δ_{SOTA}	↑ **3.36**	↑ **1.57**	↑ **5.85**	↑ **8.06**	↑ **0.44**	↑ **10.46**

Table 3: Performance comparison of different methods on Twitter-GMNER and Twitter-FMNERG datasets. **Bold** represents the optimal result, and underlined represents the suboptimal result.

et al., 2018), BERT (Devlin et al., 2019), and sequence generation methods like BARTNER (Lewis et al., 2020). (2) Two-stage models: Extract textual entities using MNER models first, then ground them to visual regions using object detectors. These methods include UMT(Yu et al., 2020), UMGF(Zhang et al., 2021a), ITA(Wang et al., 2022), BARTMNER (Lewis et al., 2020), RiVEG(Li et al., 2024a), GEM(Wang et al., 2024). (3) Unified-Generative models: Simultaneously predict textual entities and their visual grounding through end-to-end generation. These methods utilize generative architectures to capture cross-modal dependencies and jointly extract entity triplets, including H-Index(Yu et al., 2023), TIGER(Wang et al., 2023a), SCANNER (Ok et al., 2024), and MQSPN(Tang et al., 2025).

4.5 Main result

The performance comparison of our method and the baselines is detailed in Table 3. We have the following observations: (1) Our method achieves the best performance in all tasks, surpassing previous approaches with improvements of 3.36% on the GMNER task and 8.06% on the FMNERG task. (2) In the MNER task, our method achieves significant improvements in both coarse-grained and fine-grained MNER. This underscores the effectiveness of our Knowledge Enhancement Agent and Entity Correction Agent in enhancing MNER accuracy and resolving semantic ambiguity. (3) The most significant improvement is in the EEG task, with a substantial increase of 10.46% in the fine-grained EEG performance. This indicates that our Entity Reasoning Grounding Agent can augment semantic information for brief entity names through multi-step Chain-of-Thought reasoning and locate precise visual region coordinates for each entity via reasoning.

4.6 Ablation Analysis

Method	GMNER	FMNERG
w/o KEA	67.63	57.50
w/o ECA	69.23	58.57
w/o SFT	64.54	50.26
w/o GRPO	68.89	56.46
MAKAR	71.88	60.54

Table 4: The ablation study results in MNERG task. "w/o" indicates the removal of the corresponding module.

To evaluate the impact of each module in our method, we conduct ablation studies on two GMNER tasks. According to the results shown in Table 4, MAKAR outperforms all variants. The performance drop for w/o KEA and w/o ECA demonstrates the importance of auxiliary knowledge to resolve semantic ambiguity in entity type and boundary prediction. Meanwhile, w/o SFT variant underperforms notably because format rewards only ensure the output format is correct, not the content

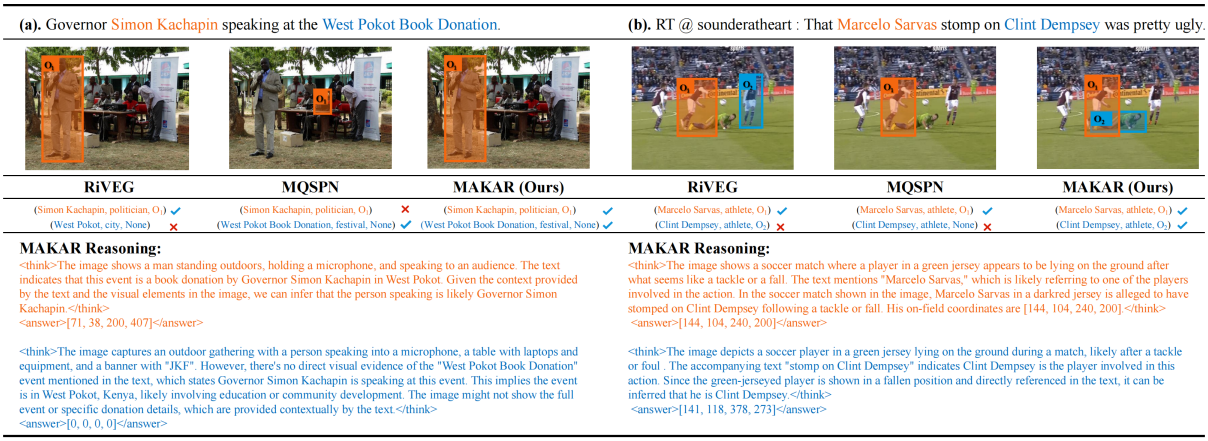

Figure 6: Prediction comparison on two test samples. ✓ and × denote correct and incorrect predictions.

of the ⟨think⟩ process. This often results in empty or irrelevant think outputs during GRPO training, emphasizing the critical role of SFT in initializing the model with basic reasoning patterns. Performance also decreases for w/o GRPO, demonstrating the effectiveness of combining SFT Cold-Start with GRPO-based reinforcement learning. While SFT provides a strong initial foundation, GRPO enhances the model's ability to handle complex cases, resulting in optimal performance. Overall, these ablation results highlight the complementary and effective roles of the modules in MAKAR.

4.7 Further Analysis

In this section, we further explore the impact of different MLLMs on the MNER tasks and the influence of different training methods on the EEG tasks. More discussions are in Appendix A and D.

4.7.1 Different MLLMs for Knowledge Enhancement and Entity Correction

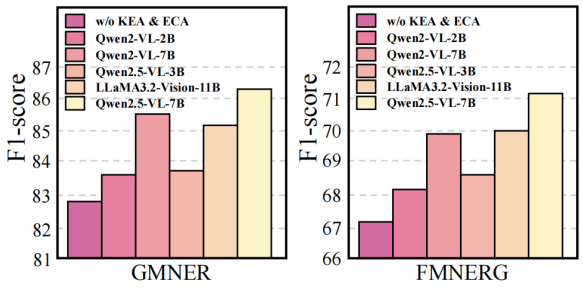

Figure 7: Performance comparison in MNER task.

As shown in Fig. 7, Qwen2.5-VL-7B outperforms other models in MNER tasks. These results highlight the effectiveness of knowledge enhancement and correction in resolving semantic ambiguity in MNER tasks.

4.7.2 Different training methods for Entity Grounding

Method	GMNER	FMNERG
Qwen2.5-VL-3B **(Raw)**	22.45	38.38
Qwen2.5-VL-3B **(SFT)**	69.88	69.09
Qwen2.5-VL-3B **(GRPO)**	45.18	58.04
Qwen2.5-VL-3B **(GRPO Cold Start)**	72.48	72.43
Qwen2.5-VL-7B **(Raw)**	60.09	60.05
Qwen2.5-VL-7B **(SFT)**	71.61	73.22
Qwen2.5-VL-7B **(GRPO)**	66.10	61.72
Qwen2.5-VL-7B **(GRPO Cold Start)**	74.64	75.66
Qwen2.5-VL-32B **(Raw)**	68.91	65.14

Table 5: Performance comparison in EEG task.

Table 5 presents the EEG performance comparison of different training methods. It shows that various fine-tuning approaches can improve entity grounding performance. Smaller models first distill reasoning processes via SFT on easy samples, and then explore optimal CoT trajectories with GRPO on hard samples. This hybrid training strategy significantly enhances the 3B model's entity grounding performance, even outperforming the 32B base model. In contrast, Qwen2.5-VL-32B performs entity grounding in a zero-shot manner without fine-tuning, and its performance is inferior to that of the smaller models (3B/7B) trained with reasoning enhancement. We further validate the effectiveness and generalizability of our two-stage training methods on additional models in Appendix A.

4.7.3 Different learning strategy for Entity Grounding

Table 6 ablates the impact of training sample difficulty in each stage of two-stage pipeline, using

SFT	RL	GMNER	FMNERG
Easy	Hard	74.64	75.66
Easy	Easy	70.47	71.57
Hard	Easy	69.55	70.46
Hard	Hard	67.82	69.04

Table 6: Performance comparison of different training sample difficulty configurations on the EEG task.

Qwen2.5-VL-7B as the base model. Our two-stage training strategy, which uses easy samples for SFT and hard samples for RL, helps the model better handle challenging scenarios while maintaining a standardized output format.

4.8 Case Study

We conduct case studies to compare our method with two baselines (RiVEG and MQSPN) on two test samples from the Twitter-FMNERG dataset. In Fig. 6(a), RIVEG fail to identify the correct entity and MQSPN fail to identify the correct bounding box of Simon Kachapin. In contrast, our method correctly identifies the correct entities and the bounding boxes. Similarly, in Fig. 6(b), we find that all methods correctly identify the two entities and types. But the two baselines fail to identify the correct bounding box of Clint Dempsey, our method accurately grounds Clint Dempsey to the visual region with the blue shadow.

5 Conclusion

In this paper, we propose MAKAR, a Multi-Agent framework based Knowledge-Augmented Reasoning for Grounded Multimodal Named Entity Recognition. Through the interaction between the Knowledge Enhancement Agent and the Entity Correction Agent, MAKAR leverages both internal and external knowledge of MLLM to address semantic ambiguity. Furthermore, the Entity Reasoning Grounding Agent enriches semantic information for each entity and performs precise reasoning grounding for each entity. Extensive experimental results demonstrated the superior performance of the MAKAR framework.

Limitations

We briefly mention some limitations of our work. First, we employ an MLLM as an implicit knowledge base to assist with textual entity recognition, which may introduce noisy information. Moreover, although our reasoning grounding method shows remarkable performance for both coarse-grained and fine-grained visual entities, it encounters challenges in identifying certain large visual regions, revealing a limitation in our coordinate generation method. Additionally, when generating the reasoning grounding process for textual entities, a relatively longer inference time is required. Considering substantial performance improvement, sacrificing a certain degree of inference speed is worthwhile. Notably, our method can be seamlessly integrated with any MLLM, and is capable of achieving better performance as MLLMs continue to advance. In the future, our research will focus on achieving optimal performance by designing lightweight modules and training them jointly.

Ethics statement

In this paper, all experimental results we provide are based on publicly available datasets and open source models. For the auxiliary knowledge and background knowledge, MAKAR generates them using Qwen and LLaMA. Therefore, we trust that all the data we use does not violate the privacy of any user.

Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by the National Key Laboratory of Science and Technology on Blind Signal Processing (No.23007522).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Chenran Cai, Qianlong Wang, Bin Liang, Bing Qin, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. In-context learning for few-shot multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2969–2979, Singapore. Association for Computational Linguistics.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction. *Preprint*, arXiv:2306.14122.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Can images help recognize entities? a study of the role of images for multimodal NER. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 87–96, Online. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 904–915. ACM.

Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Haonan Li, and 5 others. 2025. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *Preprint*, arXiv:2506.14965.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1440–1448, USA. IEEE Computer Society.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Jinyuan Li, Han Li, Di Sun, Jiahao Wang, Wenkun Zhang, Zan Wang, and Gang Pan. 2024a. LLMs as bridges: Reformulating grounded multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1302–1318, Bangkok, Thailand. Association for Computational Linguistics.

Jinyuan Li, Ziyan Li, Han Li, Jianfei Yu, Rui Xia, Di Sun, and Gang Pan. 2024b. Advancing grounded multimodal named entity recognition via llm-based reformulation and box-based segmentation. *Preprint*, arXiv:2406.07268.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, and 1 others. 2024c. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Peipei Liu, Gaosheng Wang, Hong Li, Jie Liu, Yimo Ren, Hongsong Zhu, and Limin Sun. 2022. Multi-granularity cross-modality representation learning for named entity recognition on social media. *Preprint*, arXiv:2210.14163.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. 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.

Feiping Nie, Ziheng Li, Rong Wang, and Xuelong Li. 2023. An effective and efficient algorithm for k-means clustering with new formulation. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3433–3443.

Hyunjong Ok, Taeho Kil, Sukmin Seo, and Jaeho Lee. 2024. SCANNER: Knowledge-enhanced approach for robust multi-modal named entity recognition of

unseen entities. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7725–7737, Mexico City, Mexico. Association for Computational Linguistics.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Jielong Tang, Zhenxing Wang, Ziyang Gong, Jianxing Yu, Xiangwei Zhu, and Jian Yin. 2025. Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25246–25254.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Feng Wang, Zesheng Shi, Bo Wang, Nan Wang, and Han Xiao. 2025. Readerlm-v2: Small language model for html to markdown and json. *Preprint*, arXiv:2503.01151.

Jieming Wang, Ziyan Li, Jianfei Yu, Li Yang, and Rui Xia. 2023a. Fine-grained multimodal named entity recognition and grounding with a generative framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3934–3943, New York, NY, USA. Association for Computing Machinery.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

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

Ziqi Wang, Chen Zhu, Zhi Zheng, Xinhang Li, Tong Xu, Yongyi He, Qi Liu, Ying Yu, and Enhong Chen. 2024. Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3211–3226, Miami, Florida, USA. Association for Computational Linguistics.

Junjie Wu, Chen Gong, Ziqiang Cao, and Guohong Fu. 2023. Mcg-mner: A multi-granularity cross-modality generative framework for multimodal ner with instruction. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3209–3218, New York, NY, USA. Association for Computing Machinery.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. 2025. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755, Abu Dhabi, UAE. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

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.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

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

Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Different training methods for Entity Grounding

Method	GMNER	FMNERG
Qwen2-VL-2B **(Raw)**	55.76	55.56
Qwen2-VL-2B **(SFT)**	56.96	57.92
Qwen2-VL-2B **(GRPO)**	55.93	56.72
Qwen2-VL-2B **(Ours, GRPO Cold Start)**	58.83	58.75
Qwen2-VL-7B **(Raw)**	58.93	58.83
Qwen2-VL-7B **(SFT)**	64.84	65.51
Qwen2-VL-7B **(GRPO)**	61.02	62.59
Qwen2-VL-7B **(Ours, GRPO Cold Start)**	66.26	65.99
Qwen2.5-VL-3B **(Raw)**	22.45	38.38
Qwen2.5-VL-3B **(SFT)**	69.88	69.09
Qwen2.5-VL-3B **(GRPO)**	45.18	58.04
Qwen2.5-VL-3B **(Ours, GRPO Cold Start)**	72.48	72.43
Qwen2.5-VL-7B **(Raw)**	60.09	60.05
Qwen2.5-VL-7B **(SFT)**	71.61	73.22
Qwen2.5-VL-7B **(GRPO)**	66.10	61.72
Qwen2.5-VL-7B **(Ours, GRPO Cold Start)**	74.64	75.66
MiMo-VL-7B **(Raw)**	64.29	63.47
MiMo-VL-7B **(SFT)**	73.46	74.09
MiMo-VL-7B **(GRPO)**	70.68	71.81
MiMo-VL-7B **(Ours, GRPO Cold Start)**	75.85	76.15
Qwen2.5-VL-32B **(Raw)**	68.91	65.14

Table 7: Performance comparison in EEG task.

Table 7 presents the EEG performance comparison of different training methods. Our GRPO Cold Start two-stage training method consistently achieves superior results across all base models (Qwen-VL (Bai et al., 2025) and MiMo-VL (Team et al., 2025)), demonstrating its broad applicability and effectiveness.

B Representative points in cluster

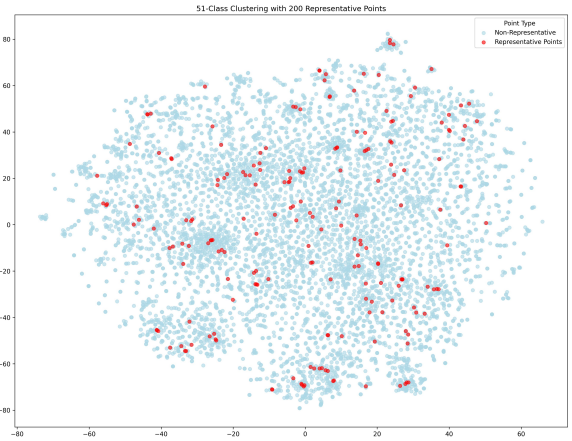

Figure 8: 200 representative points in 51 class cluster.

Fig. 8 displays the clustering results of 200 representative points across 51 entity types. To enhance knowledge for named entity recognition, we employ K-Means clustering based on entity types and multimodal representations. The 200 representative samples, shown as red dots in the figure, are selected to form the Query Set. This query set provides auxiliary examples to improve the Knowledge Enhancement Agent's performance by clarifying the semantic boundaries of different categories.

C Entity Correction

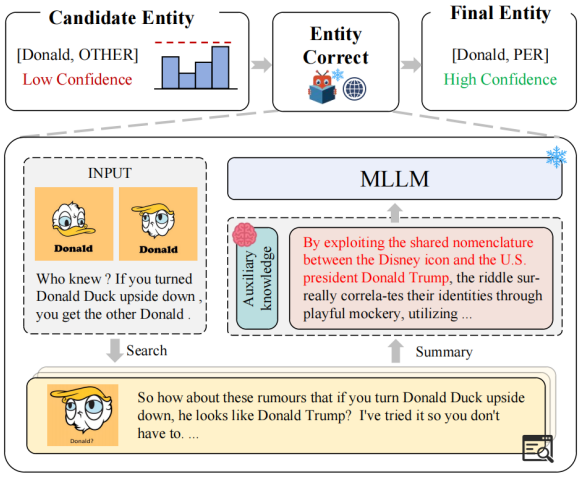

Figure 9: Example of Entity Correction.

In this section, we present an example (as shown in Fig. 9) illustrating the correction process for

low-confidence entities. Initially, the Knowledge Enhancement Agent identifies [Donald, OTHER] as a low-confidence candidate entity. To refine this entity, the Entity Correction Agent is employed. First, leveraging Qwen2.5-VL-7B, we generate a search query based on the input text and invoke the Bing search engine to retrieve three most relevant web pages (Li et al., 2024c). Subsequently, ReaderLM-v2 (Wang et al., 2025) is utilized to convert these web pages into JSON format and extract information related to the original text and the candidate entity, which is then summarized. Finally, by integrating the auxiliary knowledge from KEA and the summarized background knowledge, the Entity Correction Agent corrects the boundaries and type of the low-confidence entity, resulting in the refined entity [Donald, PER]. This example demonstrates how the Entity Correction Agent effectively addresses semantic ambiguity by dynamically incorporating external knowledge.

D Hyperparameter Sensitivity Analysis

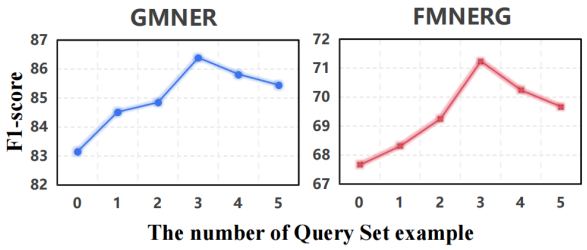

Figure 10: Analysis of the number of Query Set example on MNER task.

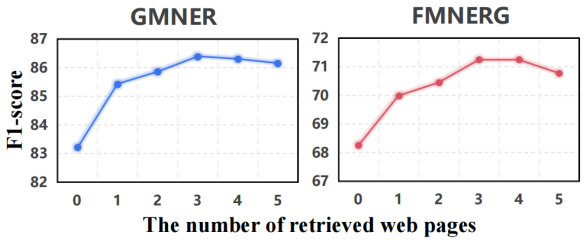

Figure 11: Analysis of the number of retrieved web pages on MNER task.

As shown in Fig. 10, we conducted an analysis on the number of Query Set examples in the MNER task. It can be observed that when the number of Query Set examples is 3, the MNER performance reaches its peak for both the GMNER and FMNERG tasks. This indicates that increasing the number of Query Set examples to three significantly enhances the model's performance. However, adding more than three examples begins to reduce performance improvement. This might be because too many examples introduce noise or conflicting information in the model's learning process.

As shown in Fig. 11, retrieving relevant web content via web search effectively corrects low-confidence entity errors. The optimal performance is achieved when retrieving 3 web pages. Moreover, we observe that relying solely on the top-1 result often fails to maximize improvement, as it may only contain the original text without introducing additional information for error correction.

E Bad case analysis

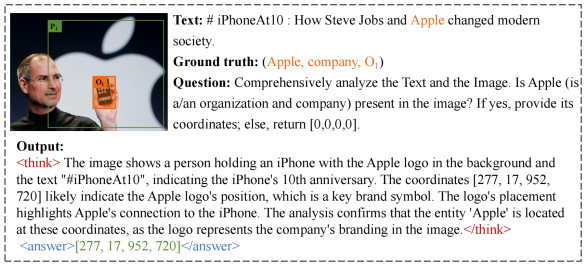

Figure 12: P_1 represents the MAKAR predicted box, and O_1 represents the ground truth.

Our MAKAR framework has limitations in grounding in some cases. As shown in Fig. 12, it localizes the Apple logo (P_1) but the ground truth is the iPhone (O_1). This deviation can not be effectively corrected by multi-step reasoning. We will address these bias issues in future research.

F FMNERG Dataset

Both of Twitter-GMNER and Twitter-FMNERG are built based on two publicly MNER Twitter datasets, i.e., Twitter-2015 (Zhang et al., 2018; Wu et al., 2023; Chen et al., 2021) and Twitter-2017 (Lu et al., 2018; Cai et al., 2023; Wang et al., 2023b). Twitter-GMNER focuses on extracting four coarse-grained types of textual entities from text-image pairs. Twitter-FMNERG, built upon GMNER, expands to 8 coarse-grained and 51 fine-grained entity types, as shown in Table 8.

G More Reasoning Grounding Example

As illustrated in Fig. 13, we compare our method with the baseline approaches on more examples. In Fig. 13(a), both baseline models misclassify users' social profile avatars as their actual identities, while

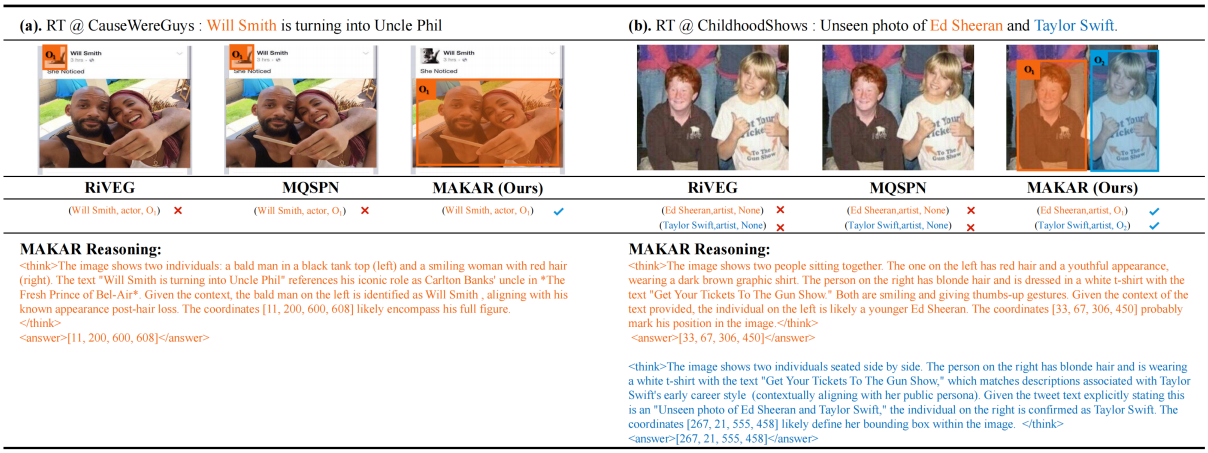

Figure 13: Prediction comparison on two test samples. ✓ and × denote correct and incorrect predictions.

our method successfully distinguishes between real individuals and their avatars. In addition, Fig. 13(b) demonstrates that baseline methods rely on current information about individuals, failing to match childhood photos of artists. However, our approach employs a reasoning grounding strategy to accurately identify these childhood images by reasoning about contextual cues. This highlights our model's ability to adapt to diverse scenarios.

H Prompt template

We present the prompt template for various instructions used at different stages of our process. In the prompt, the blue text indicates elements that should be changed according to the sample, while the black text remains constant. Notably, our method can be seamlessly integrated with any MLLM, and is capable of achieving better performance as MLLMs continue to advance.

Prompt Template for Generating Chain-of-Thought Reasoning Process

Image: {Input Image}
Text: {Input Text}
Question: Given Entity {Named Entity}, belongs to a/an {Entity Type}, based on the visual information in the image and the semantic information in the text, it is inferred that the Position of the Entity {Named Entity} in the image is {Entity Position}.

Notes: Please comprehensively consider the object features, positional relationships, descriptions in the text and possible semantic associations in this area of the image, and provide reasonable and detailed inference results.

Answer:

Prompt Template for Entity Reasoning Grounding

Image: {Input Image}
Text: {Input Text}
Question: Comprehensively analyze the text and the image. Is {Entity Name} (belongs to {Coarse-grained Entity Type} and {Fine-grained Entity Type}) present in the image? If yes, provide its coordinates; else return $[0, 0, 0, 0]$.

Note: Please ignore text content in the image (e.g., titles, tags, text boxes, etc.) and focus only on non-text visual elements (e.g., people, objects, logos, etc.).

Answer:

Prompt Template for Refining and Filtering CoT Reasoning Process

Reasoning Process: {CoT Reasoning Process}
Question: Please optimize the Chain-of-Thought reasoning analysis following these requirements.

Optimization Requirements:

- Validate the coherence of logical reasoning.
- Verify the alignment between step-by-step analysis and multimodal evidence.
- Correct any flawed assumptions or factual inaccuracies.

Output Format: <answer> Revised reasoning process. </answer>

Output:

Prompt Template for Web Search

Candidate Entity: {Low-Confidence Candidate Entity}
Input Text: {Input Text}
Task: Given the candidate entity, please use web search tools to find the 3 most relevant web pages based on the input text. The search should focus on gathering background information that can help correct the entity's boundary and type.
Notes: The search results should be relevant to the candidate entity and the context provided in the input text. The goal is to retrieve information that can clarify the entity's boundaries and type, especially in cases where long-tail distributions may have caused initial errors.

Recent chat:
Please reason step by step.
1. The "thinking" phase: You need to find background information to correct the entity's boundary and type.
2. Tool Invocation Phase: You will use the search tool to find relevant web pages.
If You think a tool is needed, You will respond with a JSON object:
{ "tool": "<tool_name>", "params": "<parameters>" }
<tool_name> only has 'search' and 'summary'. You don't generate anything else.

Example:
if You need to search, You will respond with something like:
{"tool": "search", "params": "Input Text"}
if You need to summary these search results, You will respond with something like:
{"tool": "summary", "params": "search results"}

Output:

Prompt Template for Correcting Low-Confidence Entity

Background Knowledge: {Background Knowledge from Web Search}
KEA Results: {Results from Knowledge Enhancement Agent}
Input Image: {Input Image}
Input Text: {Input Text}
Task: Based on the background knowledge retrieved from web search, along with results from KEA, the input image, and text, please refine the entity boundaries and types. Summarize the key information that can help correct the entities and provide a clear and concise explanation of the corrections.

Notes: The goal is to use this background knowledge to address semantic ambiguity and improve the accuracy of named entity recognition.

Output:

Prompt Template for Directly Extracting Textual Entity

Question: Here are some content that people post on Twitter, and these content are composed of original text and image of the original text. Please note that the text and image here may or may not be relevant, so make your own judgment. Please follow the data annotation style and method reflected in the example I provided, comprehensively analyze the image and the original text, determine which named entities and their corresponding types are included in the original text. There will only be 4 types of entities: ['LOC', 'MISC', 'ORG', 'PER']. Make the answer format like: ['entity name1', 'entity type1'],['entity name2', 'entity type2'].....

Note:

- 1. Only analyze entities in the 'Text', not in 'Image descriptions'.
- 2. Don't change the writting style and format of entity names in original Text.
- 3. The words beginning with @ sign are not counted.

Entity Definitions:

- PER (Person): People's name and fictional character.
- LOC (Location): Country, city, town continent by geographical location.
- ORG (Organization): Include club, company, government party, school government, and news organization.
- MISC (Miscellaneous): Named entities that do not belong to the types of LOC, PER, and ORG, including but not limited to event, concept, product, natural phenomenon, etc..

Text: Podcast : Cavs - Warriors Game 3 recap # Cavaliers # NBA
Image descriptions: A player in a yellow jersey shoots the basketball while being defended by an opposing player in a dark jersey.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text?
Answer: [Cavs, ORG], [Warriors, ORG], [NBA, ORG], [Cavaliers, ORG]

Text: Kevin Durant has more points (23) than the Splash Bros combined (22) .
Image descriptions: A basketball player in a white Thunder jersey with the number 35 dribbles the ball during a game, surrounded by an energetic crowd in the background.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text?
Answer: [Kevin Durant, PER], [Splash Bros, PER]

Text: Russell Westbrook on Stephen Curry: " He's not nothing I haven't seen" OKC v Warriors
Image descriptions: A basketball player in a white jersey passes the ball while being closely guarded by two opponents in blue jerseys, with a packed crowd watching from the stands.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text?
Answer: [Russell Westbrook, PER], [Stephen Curry, PER], [OKC, ORG],[Warriors, ORG]

Text: My son was excited about the win # DubNation # NBAFinals # customjersey # GoldenStateWarriors
Image: {Input Image}
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text?
Answer:

Prompt Template for Providing Auxiliary Knowledge to Coarse-Grained Entity

Question: Here are some content that people post on Twitter, and these content are composed of original text and image of the original text. Please note that the text and image here may or may not be relevant, so make your own judgment. Please follow the data annotation style and method reflected in the example I provided, comprehensively analyze the image and the original text, determine which named entities and their corresponding types are included in the original text, and explain the reason for your judgment. There will only be 4 types of entities: ['LOC', 'MISC', 'ORG', 'PER'].

Note:

- 1. Only analyze entities in the 'Text', not in 'Image descriptions'.
- 2. Don't change the writting style and format of entity names in original Text.
- 3. The words beginning with @ sign are not counted.

Entity Definitions:

- PER (Person): People's name and fictional character.
- LOC (Location): Country, city, town continent by geographical location.
- ORG (Organization): Include club, company, government party, school government, and news organization.
- MISC (Miscellaneous): Named entities that do not belong to the types of LOC, PER, and ORG, including but not limited to event, concept, product, natural phenomenon, etc..

Text: Podcast : Cavs - Warriors Game 3 recap # Cavaliers # NBA
Image descriptions: A player in a yellow jersey shoots the basketball while being defended by an opposing player in a dark jersey.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer: Named entities: 1. Cavs (Cleveland Cavaliers, basketball team) 2. Warriors (Golden State Warriors, basketball team) 3. NBA (league/organization) 4.Cavaliers (Cleveland Cavaliers, basketball team) Reasoning: The text mentions "Cavs" and "Warriors" which refer to the Cleveland Cavaliers and Golden State Warriors respectively, both of which are basketball teams. Additionally, "NBA" is mentioned, which stands for the National Basketball Association, a professional basketball league in the United States. The image description does not provide any additional information about named entities beyond what is mentioned in the text.

Text: Kevin Durant has more points (23) than the Splash Bros combined (22) .
Image descriptions: A basketball player in a white Thunder jersey with the number 35 dribbles the ball during a game, surrounded by an energetic crowd in the background.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer: Named entities: 1. Kevin Durant (person/player), 2. Splash Bros (team/organization) Reasoning: The sentence refers to Kevin Durant, a professional basketball player, and the "Splash Bros," which is a nickname for the Golden State Warriors' starting lineup consisting of Stephen Curry, Klay Thompson, Draymond Green, and Andre Iguodala. The sentence compares Durant's points scored (23) to the combined points scored by the Splash Bros (22). The mention of the Splash Bros implies that they are a team, specifically the Golden State Warriors.

Text: Russell Westbrook on Stephen Curry: " He's not nothing I haven't seen" OKC v Warriors
Image descriptions: A basketball player in a white jersey passes the ball while being closely guarded by two opponents in blue jerseys, with a packed crowd watching from the stands.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer: Named entities: 1. Russell Westbrook (person/player) 2. Stephen Curry (person/player) 3. OKC (team/franchise) 4. Warriors (team/franchise) Reasoning: The sentence mentions Russell Westbrook and Stephen Curry, both of whom are professional basketball players. The text also refers to "OKC," which stands for Oklahoma City Thunder, and "Warriors," which refers to the Golden State Warriors, both of which are professional basketball teams. The sentence suggests a comparison between Westbrook's opinion of Curry and their upcoming match between the OKC Thunder and the Warriors.

Text: My son was excited about the win # DubNation # NBAFinals # customjersey # GoldenStateWarriors
Image: {Input Image}
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer:

Prompt Template for Providing Auxiliary Knowledge to Fine-Grained Entity

Question: Here are some content that people post on Twitter, and these content are composed of original text and image of the original text. Please note that the text and image here may or may not be relevant, so make your own judgment. Please follow the data annotation style and method reflected in the example I provided, comprehensively analyze the image and the original text, determine which named entities and their corresponding types are included in the original text, and explain the reason for your judgment. There will be 51 types of entities organized into 8 coarse categories.

Note:

- 1. Only analyze entities in the 'Text', not in 'Image descriptions'.
- 2. Preserve original writing style/format of entity names in Text.
- 3. Ignore words beginning with @ symbols.
- 4. Use fine-grained types (e.g., product-brand_name_products)

Entity Definitions:

- **[LOCATION]**:
  - location-city: Cities (e.g., "New York", "Tokyo")
  - location-country: Sovereign states (e.g., "Canada", "Australia")
  - location-state: Administrative regions (e.g., "California", "Queensland", "Taiwan")
  - location-continent: Continents (e.g., "Africa", "Europe")
  - location-other: General locations (e.g., "Central Park", "Mount Everest")
  - location-park: Public parks (e.g., "Hyde Park", "Yosemite")
  - location-road: Streets/highways (e.g., "Route 66", "Oxford Street")
- **[BUILDING]**:
  - building-other: Generic structures (e.g., "Empire State Building")
  - building-cultural_place: Museums/libraries (e.g., "Louvre Museum")
  - building-entertainment_place: Theaters/cinemas (e.g., "Madison Square Garden")
  - building-sports_facility: Stadiums/arenas (e.g., "Wembley Stadium")
- **[ORGANIZATION]**:
  - organization-company: Businesses (e.g., "Apple", "Toyota")
  - organization-educational_institution: Schools/universities (e.g., "Harvard University")
  - organization-band: Music groups (e.g., "Coldplay", "BTS")
  - organization-government_agency: Government bodies (e.g., "FBI", "NHS")
  - organization-news_agency: Media outlets (e.g., "BBC", "Reuters")
  - organization_other: General organizations
  - organization-political_party: Political groups (e.g., "Republican Party")
  - organization-social_organization: NGOs/clubs (e.g., "Red Cross")
  - organization-sports_league: Athletic leagues (e.g., "NBA", "Premier League")
  - organization-sports_team: Sports clubs (e.g., "LA Lakers", "Manchester United")
- **[PERSON]**:
  - person-politician: Government officials (e.g., "Joe Biden")
  - person-musician: Singers/instrumentalists (e.g., "Taylor Swift")
  - person-actor: Film/TV performers (e.g., "Leonardo DiCaprio")
  - person-artist: Visual artists (e.g., "Picasso")
  - person-athlete: Sports professionals (e.g., "Serena Williams")
  - person-author: Writers (e.g., "J.K. Rowling")
  - person-businessman: Corporate leaders (e.g., "Elon Musk")
  - person-character: Fictional characters (e.g., "Harry Potter")
  - person-coach: Sports trainers (e.g., "Gregg Popovich")
  - person-common_person: Ordinary individuals
  - person-director: Film directors (e.g., "Christopher Nolan")
  - person-intellectual: Scholars/thinkers (e.g., "Albert Einstein")
  - person-journalist: Reporters (e.g., "Anderson Cooper")
  - person_other: Miscellaneous person references

- **[OTHER]**:
  - other-animal: Animal names/species (e.g., "African Elephant")
  - other-award: Prizes/honors (e.g., "Nobel Prize")
  - other-medical_thing: Medical terms (e.g., "MRI machine")
  - other-website: Web domains (e.g., "Wikipedia.org")
  - other-ordinance: Laws/regulations (e.g., "GDPR")
- **[ART]**:
  - art-art_other: General art references
  - art-film_and_television_works: Movies/TV shows (e.g., "Stranger Things")
  - art-magazine: Publications (e.g., "Vogue", "Time")
  - art-music: Song/album titles (e.g., "Thriller")
  - art-written_work: Books/articles (e.g., "1984")
- **[EVENT]**:
  - event-event_other: General events
  - event-festival: Cultural festivals (e.g., "Coachella")
  - event-sports_event: Athletic competitions (e.g., "Olympics")
- **[PRODUCT]**:
  - product-brand_name_products: Branded goods (e.g., "iPhone 15")
  - product-game: Video/board games (e.g., "Minecraft")
  - product-product_other: General products
  - product-software: Applications (e.g., "Photoshop")

Text: Podcast : Cavs - Warriors Game 3 recap # Cavaliers # NBA
Image descriptions: A player in a yellow jersey shoots the basketball while being defended by an opposing player in a dark jersey.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer: Named entities: 1. Cavs (Cleveland Cavaliers, basketball team, organization-sports_team) 2. Warriors (Golden State Warriors, basketball team, organization-sports_team) 3. NBA (league/organization ,organization-sports_league) 4.Cavaliers (Cleveland Cavaliers, basketball team, organization-sports_team) Reasoning: The text mentions "Cavs" and "Warriors" which refer to the Cleveland Cavaliers and Golden State Warriors respectively, both of which are basketball teams. Additionally, "NBA" is mentioned, which stands for the National Basketball Association, a professional basketball league in the United States.

Text: Kevin Durant has more points (23) than the Splash Bros combined (22) .
Image descriptions: A basketball player in a white Thunder jersey with the number 35 dribbles the ball during a game, surrounded by an energetic crowd in the background.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer: Named entities: 1. Kevin Durant (person/player,person-athlete), 2. Splash Bros (person/player,person-athlete) Reasoning: The sentence refers to Kevin Durant, a professional basketball player, and the "Splash Bros," which is a nickname for the Golden State Warriors' starting lineup consisting of Stephen Curry, Klay Thompson, Draymond Green, and Andre Iguodala. The sentence compares Durant's points scored (23) to the combined points scored by the Splash Bros (22). The mention of the Splash Bros implies that they are a team, specifically the Golden State Warriors.

Text: Russell Westbrook on Stephen Curry: " He's not nothing I haven't seen" OKC v Warriors
Image descriptions: A basketball player in a white jersey passes the ball while being closely guarded by two opponents in blue jerseys, with a packed crowd watching from the stands.
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer: Named entities: 1. Russell Westbrook (person/player,person-athlete) 2. Stephen Curry (person/player,person-athlete) 3. OKC (team/franchise,organization-sports_team) 4. Warriors (team/franchise,organization-sports_team) Reasoning: The sentence mentions Russell Westbrook and Stephen Curry, both of whom are professional basketball players. The text also refers to "OKC," which stands for Oklahoma City Thunder, and "Warriors," which refers to the Golden State Warriors, both of which are professional basketball teams. The sentence suggests a comparison between Westbrook's opinion of Curry and their upcoming match between the OKC Thunder and the Warriors.

Text: My son was excited about the win # DubNation # NBAFinals # customjersey # GoldenStateWarriors
Image: {Input Image}
Question: Comprehensively analyze the Text and the Image description, which named entities and their corresponding types are included in the Text? explain the reason for your judgment.
Answer:

Fine-Grained	Coarse-Grained	Definition
city	location	Cities (e.g., "New York", "Tokyo")
country	location	Sovereign states (e.g., "Canada", "Australia")
state	location	Administrative regions (e.g., "California", "Queensland")
continent	location	Continents (e.g., "Africa", "Europe")
park	location	Public parks (e.g., "Hyde Park", "Yosemite")
road	location	Streets/highways (e.g., "Route 66", "Oxford Street")
other	location	General locations (e.g., "Central Park", "Mount Everest")
cultural place	building	Museums/libraries (e.g., "Louvre Museum")
entertainment place	building	Theaters/cinemas (e.g., "Madison Square Garden")
sports facility	building	Stadiums/arenas (e.g., "Wembley Stadium")
other	building	Generic structures (e.g., "Empire State Building")
company	organization	Businesses (e.g., "Apple", "Toyota")
educational institution	organization	Schools/universities (e.g., "Harvard University")
band	organization	Music groups (e.g., "Coldplay", "BTS")
government agency	organization	Government bodies (e.g., "FBI", "NHS")
news agency	organization	Media outlets (e.g., "BBC", "Reuters")
political party	organization	Political groups (e.g., "Republican Party")
social organization	organization	NGOs/clubs (e.g., "Red Cross")
sports league	organization	Athletic leagues (e.g., "NBA", "Premier League")
sports team	organization	Sports clubs (e.g., "LA Lakers", "Manchester United")
other	organization	General organizations
politician	person	Government officials (e.g., "Joe Biden")
musician	person	Singers/instrumentalists (e.g., "Taylor Swift")
actor	person	Film/TV performers (e.g., "Leonardo DiCaprio")
artist	person	Visual artists (e.g., "Picasso")
athlete	person	Sports professionals (e.g., "Serena Williams")
author	person	Writers (e.g., "J.K. Rowling")
businessman	person	Corporate leaders (e.g., "Elon Musk")
character	person	Fictional characters (e.g., "Harry Potter")
coach	person	Sports trainers (e.g., "Gregg Popovich")
common person	person	Ordinary individuals
director	person	Film directors (e.g., "Christopher Nolan")
intellectual	person	Scholars/thinkers (e.g., "Albert Einstein")
journalist	person	Reporters (e.g., "Anderson Cooper")
other	person	Miscellaneous person references
art other	art	General art references
film and television works	art	Movies/TV shows (e.g., "Stranger Things")
magazine	art	Publications (e.g., "Vogue", "Time")
music	art	Song/album titles (e.g., "Thriller")
written work	art	Books/articles (e.g., "1984")
event other	event	General events
festival	event	Cultural festivals (e.g., "Coachella")
sports event	event	Athletic competitions (e.g., "Olympics")
brand name products	product	Branded goods (e.g., "iPhone 15")
game	product	Video/board games (e.g., "Minecraft")
product other	product	General products
software	product	Applications (e.g., "Photoshop")
animal	other	Animal names/species (e.g., "African Elephant")
award	other	Prizes/honors (e.g., "Nobel Prize")
medical thing	other	Medical terms (e.g., "MRI machine")
website	other	Web domains (e.g., "Wikipedia.org")
ordinance	other	Laws/regulations (e.g., "GDPR")

Table 8: Entity type and definition of the FMNERG dataset.