ISR: Self-Refining Referring Expressions for Entity Grounding

Zhuocheng Yu[1], Bingchan Zhao[2], Yifan Song[1], Sujian Li[1]*, Zhonghui He[3]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2]University of Washington
[3]Department of PE, Peking University
yzc@stu.pku.edu.cn, lisujian@pku.edu.cn

Abstract

Entity grounding, a crucial task in constructing multimodal knowledge graphs, aims to align entities from knowledge graphs with their corresponding images. Unlike conventional visual grounding tasks that use referring expressions (REs) as inputs, entity grounding relies solely on entity names and types, presenting a significant challenge. To address this, we introduce a novel **I**terative **S**elf-**R**efinement (**ISR**) scheme to enhance the multimodal large language model's capability to generate high quality REs for the given entities as explicit contextual clues. This training scheme, inspired by human learning dynamics and human annotation processes, enables the MLLM to iteratively generate and refine REs by learning from successes and failures, guided by outcome rewards from a visual grounding model. This iterative cycle of self-refinement avoids overfitting to fixed annotations and fosters continued improvement in referring expression generation. Extensive experiments demonstrate that our methods surpasses other methods in entity grounding, highlighting its effectiveness, robustness and potential for broader applications[1].

1 Introduction

As a crucial subtask in the construction of multimodal knowledge graphs, entity grounding (EG) task aims to ground entities in knowledge graphs to their corresponding multimodal data, especially images (Zhu et al., 2022). Following recent research (Wang et al., 2023; Yu et al., 2023; Tang et al., 2024), we focus on a fine-grained entity grounding task: given an entity and an image, the task requires the model to first determine whether the entity is present in the image (i.e., whether it can be grounded) and, for groundable entities, to further provide their corresponding visual regions (bounding boxes).

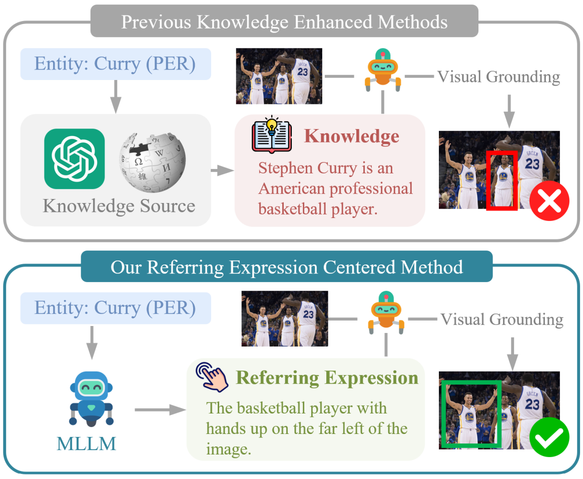
Previous Knowledge Enhanced Methods
Entity: Curry (PER)
Visual Grounding
Knowledge Source
Knowledge
Stephen Curry is an American professional basketball player.
Our Referring Expression Centered Method
Entity: Curry (PER)
Visual Grounding
MLLM
Referring Expression
The basketball player with hands up on the far left of the image.

Figure 1: The comparison between existing knowledge-enhanced entity grounding method and our referring expression-centered method.

Unlike conventional visual grounding (i.e., referring expression comprehension) tasks (Mao et al., 2016; Yu et al., 2016), the input of entity grounding task only consists of an entity and an image, without *referring expression* (RE), which is an unambiguous text description of exactly one object or region in the image. The absence of REs can pose significant obstacles for entity grounding, such as ambiguity in entity identification when the image is complex or contains multiple similar objects. For example, as illustrated in Figure 1, without an RE, the grounding model may face challenges in accurately identifying "Stephen Curry" among three visually similar basketball players. In light of the success of methods (Deng et al., 2021; Wang et al., 2022a; Li et al., 2023c) in visual grounding, it is intuitive to incorporate REs into the entity grounding task as explicit contextual clues for the model to ground entities.

So far, acquiring accurate REs typically requires manual annotations (Kazemzadeh et al., 2014;

*Corresponding author.
[1] Code & Data: https://github.com/Zhuocheng0579/ISR.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30702–30714
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Plummer et al., 2015), which is labor-intensive and impractical for the entity grounding task. Existing automated referring expression generation methods (Yu et al., 2016; Mao et al., 2016; Sun et al., 2022; Ye et al., 2023) need to take visual regions (i.e., bouding boxes) as input, limiting their applicability to the entity grounding task as they lack the capability to directly generate REs for the given entities without bounding boxes. Recent efforts utilize the knowledge related to the given entity as referring expression, acquiring knowledge from Wikipedia (Ok et al., 2024) or large language models (Li et al., 2024b). Nevertheless, the knowledge employed in these methods remains limited to factual descriptions and fails to adequately integrate visual information, making it still challenging for the grounding model to provide correct answer, as illustrated in Figure 1 (a).

To effectively incorporate REs into the entity grounding task, we train a multimodal large language model (MLLM, Li et al., 2024a; Wang et al., 2024b; Chen et al., 2024) to generate high quality REs for the given entities, then utilize a visual grounding model to give the bounding box predictions according to the REs, as illustrated in Figure 1 (b). The focus is on how to enhance the capability of MLLMs to generate accurate and unambiguous REs for the given entities. A straight-forward approach for training the MLLM is to directly perform supervised fine-tuning (SFT) on expert REs annotated by human or advanced AI models (e.g. GPT-4o). However, standard SFT methods overly rely on the fixed set of expert-annotated REs, which can lead to overfitting and hinder continued improvement during training.

As we know, humans usually master a new skill in continuous exploration and experimentation, learning from feedback to progressively enhance their proficiency. Building on this observation, we propose a novel **I**terative **S**elf-**R**efinement (**ISR**) training scheme for referring expression generation, which encourages the MLLM to explore valid ways of describing the target entity during training and learn from both successes and failures. To obtain feedback for the model during training, we draw inspiration from ReferitGame (Kazemzadeh et al., 2014), which is widely adopted in the process of human referring expression annotation. In this two-player game, Player 1 writes RE for the target object, and Player 2 is asked to click on the region corresponding to the RE written by Player 1. The correctness of Player 2 serves as feedback on the quality of the written RE. Analogously, at each iteration of the training loop, we let the current MLLM act as Player 1, generating multiple REs for each given entity. We employ a visual grounding model as Player 2, providing bounding box predictions according to the generated REs. The correctness then serves as an outcome reward to judge the quality of the REs. Based on the outcome rewards, we construct the preference data to perform Direct Preference Optimization (DPO, Rafailov et al., 2024) training, thereby enhancing the MLLM's capability.

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

- As far as we know, we are the first to incorporate referring expression generation into the entity grounding task, enhancing its process by providing descriptive textual clues.
- We introduce the ISR training scheme, that encourages the MLLM to explore diverse ways of generating REs while learning from both correct and incorrect outcomes, without over-reliance on a fixed set of expert annotations.
- Through extensive experiments, we demonstrate that our methods significantly outperform previous methods for entity grounding.

2 Related Work

Entity Grounding As an integral part of constructing multimodal knowledge graphs, entity grounding focuses on aligning entities within the knowledge graph to their corresponding multimodal information, particularly visual data. Previous methods primarily rely on online encyclopedic resources (Wang et al., 2020; Alberts et al., 2020) or search engines (Oñoro-Rubio et al., 2017; Liu et al., 2019) to retrieve images for a given entity. However, these approaches often results in noise, as the retrieved images may not contain the specified entity or may include other unrelated entities. Consequently, recent research has shifted focus toward a more fine-grained entity grounding task (Wang et al., 2023; Yu et al., 2023; Tang et al., 2024). Given an entity and an image, the task requires the model to first ascertain whether the entity is groundable and for groundable entities, to further provide their corresponding visual regions. Many previous studies (Yu et al., 2023; Ok et al., 2024) rely on object detection models (Zhang et al., 2021b; Girshick et al., 2014) to identify candidate regions, which are then matched to entities.

However, the performance of these approaches is inherently constrained by the capabilities of the object detection models. To address this limitation, Li et al. (2024b) explored the use of visual grounding models, incorporating entity-related knowledge as referring expressions. Nonetheless, since the knowledge does not directly include visual information, it poses challenges for visual grounding models to make accurate predictions.

Multimodal Large Language Models Recent work on multimodal large language models has made significant progress in integrating and processing multiple modalities, including text, images, and audio. These models aim to improve cross-modal reasoning and generation, advancing capabilities toward more general-purpose AI systems. Several approaches, such as Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), and MiniGPT (Zhu et al., 2023), have contributed to better cross-modal understanding and generation. Additionally, researchers have focused on aligning model outputs with human preferences to ensure the generated content adheres to principles of being helpful, honest, and harmless (Bai et al., 2022a). Techniques such as reinforcement learning with human feedback (RLHF) and its multimodal extensions (Zhao et al., 2023; Wang et al., 2024a; Zhang et al., 2024) have been applied to enhance reliability and reduce harmful outputs. These methodologies also help make MLLMs more interpretable and ethically robust, supporting their potential use in real-world applications.

3 Methodology

During inference phase, as depicted in Figure 1 (b), we utilize an MLLM to generate an RE for the given entity. The generated RE then serves as the input of a visual grounding model, which provides the bounding box prediction according to the RE. In this section, we first collect the expert REs for training (Section 3.2). Then, we briefly introduce the fine-tuning of the visual grounding model (Section 3.3). Finally, we elaborate on the ISR training process for enhancing the MLLM's ability to generate accurate REs (Section 3.4).

3.1 Task Formulation

Given an entity e, its type $c \in \{PER, LOC, ORG, MISC\}$, and an image I, the goal of entity grounding is to map the entity to its corresponding visual region v in the image. If e appears in I, v consists of a 4D coordinates representing the top-left and bottom-right locations of the grounded bounding box, i.e., (x_1, y_1, x_2, y_2). Otherwise, v should be $None$.

3.2 Expert Referring Expressions Collection

Define $\mathbb{E}$ as the set of all entities in the dataset, $\mathbb{E}_g$ as the subset of groundable entities and $\mathbb{E}_u$ as the subset of ungroundable entities. For each groundable entity $e \in \mathbb{E}_g$, we draw its bounding box[2] on the corresponding image I, then we prompt GPT-4o to generate the expert referring expression s. The prompt used here could be seen in Appendix A.1. For each ungroundable entity $e \in \mathbb{E}_u$, we define its expert RE as "*The entity does not appear in the image*", which we denote as s_u. Additionally, we manually check the REs generated by GPT-4o for accuracy and make corrections where needed.

3.3 Visual Grounding Model Fine-tuning

The visual grounding model takes referring expression and image as input, then predicts the visual region (bouding box) according to the referring expression. It plays two roles in our method: (1) during the training phase, it provides feedback for the REs generated by the MLLM (Section 3.4.2); (2) during inference phase, it gives the final predictions. We employ off-the-shelf visual grounding models in this work, fine-tuning them on the expert REs and ground-truth bounding boxes:

$$\mathcal{D}_{\text{VG}} = \left\{ (I, s, v)^{(i)} \middle| e^{(i)} \in \mathbb{E}_g \right\} \tag{1}$$

where v represents the ground-truth bounding box and $\mathbb{E}_g$ denotes the set of groundable entities. We train the VG model by minimizing the loss:

$$\mathcal{L}_{\text{VG}}(\theta) = -\mathbb{E}_{(I,s,v)\sim\mathcal{D}_{\text{VG}}}[\log P_\theta(v|I,s)] \tag{2}$$

where θ refers to the model parameters.

3.4 Iterative Self-Refinement of RE Generation

The complete training process of ISR consists of two stages: supervised fine-tuning initialization (Section 3.4.1) and iterative preference learning (Section 3.4.2).

[2] In cases where there are multiple ground-truth bounding boxes, we select the largest one by area.

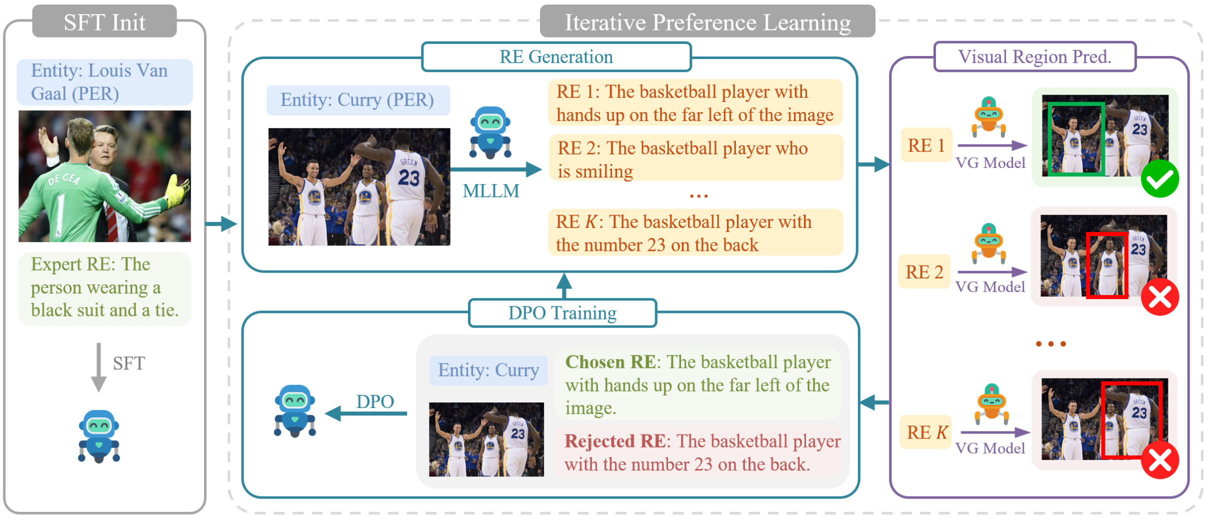
SFT Init
Entity: Louis Van Gaal (PER)
Expert RE: The person wearing a black suit and a tie.
SFT
Iterative Preference Learning
RE Generation
Entity: Curry (PER)
MLLM
RE 1: The basketball player with hands up on the far left of the image
RE 2: The basketball player who is smiling
...
RE K: The basketball player with the number 23 on the back
Visual Region Pred.
RE 1
VG Model
RE 2
VG Model
RE K
VG Model
DPO Training
DPO
Entity: Curry
Chosen RE: The basketball player with hands up on the far left of the image.
Rejected RE: The basketball player with the number 23 on the back.

Figure 2: The overall architecture of ISR. The MLLM is first fine-tuned on expert REs. Then we optimize the MLLM through iterative preference learning.

3.4.1 Supervised Fine-tuning Initialization

To equip the MLLM with a foundational ability to generate REs based on the given entities and images, we first perform supervised fine-tuning on the model. We randomly select a small subset of samples $\mathcal{D}_{\text{SFT}}$ for supervised fine-tuning, while the remaining samples, referred to as $\mathcal{D}_{\text{PL}}$, were reserved for iterative preference learning. We fine-tune the MLLM on the auto-regressive loss to get the π_0:

$$\mathcal{L}_{\text{SFT}}(\pi_\theta) = -\mathbb{E}_{(p,s)\sim\mathcal{D}_{\text{SFT}}}[\log \pi_\theta(s|p)] \quad (3)$$

where p denotes the prompt for the MLLM, consisting of instruction, entity and image (see Appendix A.2) and s is the expert referring expression.

3.4.2 Iterative Preference Learning

Starting from the supervised fine-tuned MLLM π_0, we adopt an iterative training scheme to guide the model toward generating more accurate REs.

Reward Definition Reward serves as a measure of how well the model's output aligns with human preferences or task objectives. Prior studies usually rely on reward models trained on human preference annotations (Ouyang et al., 2022; Bai et al., 2022a; Dubey et al., 2024), or advanced AI models (Bai et al., 2022b; Lee et al., 2023) to assign rewards for the model's outputs. However, these approaches demand additional cost of human labeling or calling APIs, and may not directly reflect the quality of the generated REs. Inspired by the human RE labeling process ReferitGame (Kazemzadeh et al., 2014), we adopt a more simple yet effective approach to define the reward of RE. Given a generated RE $\hat{s}$ of an entity e, (1) if e is groundable, we feed the RE and image into the VG model to get the predicted visual region:

$$\hat{v} = M_{\text{VG}}(I, \hat{s}) \quad (4)$$

where M_{VG} denotes the VG model, $\hat{s}$ represents the RE generated by the MLLM and $\hat{v}$ is the predicted visual region. Then we check the correctness of the $\hat{v}$ to assign a reward for $\hat{s}$. (2) If e is ungroundable, we just check if $\hat{s}$ is s_u (i.e., *The entity does not appear in the image*). Formally, we define the binary reward of $\hat{s}$ as:

$$r = \begin{cases} 1, & e \in \mathbb{E}_g \land \hat{v} \text{ is correct} \\ 1, & e \in \mathbb{E}_u \land \hat{s} = s_u \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where $\mathbb{E}_g$ is the set of groundable entities and $\mathbb{E}_u$ is the set of ungroundable entities.

Preference Dataset Construction At iteration t, we sample a mini-batch of prompts from $\mathcal{D}_{\text{PL}}$. For each prompt $p \in \mathcal{D}_{\text{PL}}$, we aim to construct an preference pair $s_w \succ s_l$, where s_w and s_l represents the better and worse RE. Specifically, for each each prompt p, we generate K different REs $\{\hat{s}_1, ..., \hat{s}_K\}$ by sampling from the model π_{t-1}, which is from the previous iteration:

$$\hat{s}_k \sim \pi_{t-1}(s|p), k = 1...K \quad (6)$$

We randomly select a generated RE with a reward of 0 as s_l. If none of the REs receive a reward of 0,

Algorithm 1: Iterative Self-Refinement

```
Input: Dataset $\mathcal{D}_{\text{PL}}$, max iteration number $T$,
       base MLLM $\pi_\theta$, sampling number $K$
Output: Final MLLM $\pi_{\theta*}$
Supervised fine-tuned $\pi_\theta$ to $\pi_0$
for $t = 1$ to $T$ do
    $\pi_{\text{ref}} = \pi_{t-1}$
    Sample a mini-batch of prompts from $\mathcal{D}_{\text{PL}}$
    Generate $K$ different REs for each prompt:
      $\hat{s}_k \sim \pi_{t-1}(s|p), k = 1...K$
    Calculate their rewards: $\{r_1, ..., r_K\}$
    Construct the preference dataset:
      $\mathcal{D}_t = \{(p, s_w, s_l)^{(i)}\}_{i=1}^{|\mathcal{D}_t|}$
    Update the MLLM's parameters:
      $\pi_t = \arg\min_{\pi_\theta} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$
end
```

we simply discard the prompt. A straight-forward choice for the s_w is the expert RE from the $\mathcal{D}_{\text{PL}}$ dataset. However, this may limit the diversity of the preference data and the model could be prone to overfitting to the expert REs. To mitigate this, we select s_w based on different situations:

1. All K generated REs have a reward of 0, which means that the MLLM currently struggles to generate a correct RE for the given entity. In this case, we choose the expert RE as s_w, which serves as the expert guidance for the MLLM on how to generate a correct RE.
2. If any of the K generated REs receives a reward of 1, we randomly select one of them as s_w. This encourages the MLLM to explore multiple valid ways of describing the target entity, and to learn from and reinforce its own successes, without overfitting to a fixed set of expert REs.

Finally, we get the preference dataset at iteration t:

$$\mathcal{D}_t = \left\{(p, s_w, s_l)^{(i)}\right\}_{i=1}^{|\mathcal{D}_t|} \tag{7}$$

DPO Training Direct Preference Optimization (DPO, Rafailov et al., 2024) offers a scalable, direct approach to preference learning by focusing on aligning the model's output to preferred responses, without the need for complex reinforcement learning algorithms. Given the preference dataset $\mathcal{D}_t$ at iteration t, DPO optimizes the model by training it to increase the likelihood of preferred RE s_w relative to the less preferred one s_l. We fine-tune the MLLM by minimizing the DPO loss:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = \\ - \mathbb{E}_{(p,s_w,s_l)\sim\mathcal{D}_t} \left[\log\sigma\left(\beta \log \frac{\pi_\theta(s_w|p)}{\pi_\theta(s_l|p)} - \beta \log \frac{\pi_{\text{ref}}(s_w|p)}{\pi_{\text{ref}}(s_l|p)}\right)\right] \tag{8}$$

This equation reflects the goal of maximizing the probability of generating the higher-reward expression s_w over the lower-reward expression s_l for a given prompt p. By iteratively constructing preference dataset and applying DPO at each step, the model gradually becomes more effective at generating accurate and contextually appropriate referring expressions, improving its performance over multiple rounds of preference learning.

4 Experiments

4.1 Experiment Settings

Datasets We conduct our experiments utilizing the images, entities and their bounding box annotations in Twitter-GMNER dataset (Yu et al., 2023). The dataset is built on two MNER datasets, i.e., Twitter-15 (Zhang et al., 2018) and Twitter-17 (Yu et al., 2020), with human bounding box annotations. The dataset contains four types of entities: person (PER), location (LOC), organization (ORG) and miscellaneous (MISC). We collect expert REs for the entities in the training set and use them to train our models. Please refer to Appendix B for more details.

Implementation Details We utilize Qwen2-VL-7B (Wang et al., 2024b) as the MLLM to generate REs, and OFA$_{\text{Large}}$ (Wang et al., 2022a) as the visual grounding model. The size of $\mathcal{D}_{\text{SFT}}$ is 1457 and the size of $\mathcal{D}_{\text{PL}}$ is 10000. For the SFT training of the MLLM, the batch size is 32 and the learning rate is 1e-4 with 3% warm up and a cosine scheduler. For sampling REs from the MLLM, we set the temperature to 0.8 and the number of samples K to 4, using vllm (Kwon et al., 2023) to accelerate the process of generation. For the DPO training, the batch size is 32 and the learning rate is 5e-6 with 10% warm up. We set the β in the DPO loss to 0.5. All the training of the MLLM uses AdamW optimizer (Loshchilov, 2017) and LoRA (Hu et al., 2021) with rank set to 32 and α set to 64. For the fine-tuning of the visual grounding model, we set the batch size to 4 and the learning rate to 3e-5. All experiments are conducted on 1 NVIDIA A800 80G GPU.

Baselines We compare our method with three types of baselines. (1) Object detection-based methods, including RCNN-EVG and VinVL-EVG (Yu et al., 2023). These methods first employ an object detection method (Girshick et al., 2014; Zhang et al., 2021b) to identify all candidate objects (i.e., visual regions), then choose the visual region with the highest probability. If all probabilities are lower than a threshold, the prediction region is *None*. (2) Knowledge-enhanced methods, including RiVEG (Li et al., 2024b) and its data-augmented variant, which attempt to utilize knowledge as a substitute for REs. (3) Closed-source advanced MLLMs, including GPT-4o and GPT-4o-mini, which exhibit robust text and visual analysis capabilities and have demonstrated remarkable performance across various multimodal tasks. (4) Fine-tuned Open-source MLLMs, including Qwen2-VL-7B (Wang et al., 2024b), InternVL2-8B (Chen et al., 2024) and LLaVA-NeXT-7B (Liu et al., 2024), which exhibit remarkable performance in visual grounding tasks. After fine-tuning on the entity grounding dataset (see Appendix A.3 for the prompt), they can serve as strong baselines for comparison. Except for the closed-source advanced MLLMs, all other methods have been fine-tuned on the training set of the entity grounding dataset.

Evaluation We evaluate all methods on the entities of the test set, and report the accuracy for four distinct entity types as well as the overall accuracy. For groundable entities, following previous work, we define a model prediction as correct if and only if the intersection-over-union (IoU) between the predicted bounding box and the ground-truth bounding box is greater than 0.5. For ungroundable entities, the model is required to output "None" to be considered correct.

4.2 Results

Table 1 shows the performance comparison of various methods on the entity grounding task across four entity types as well as the overall accuracy. Our proposed method achieves SOTA performance with an overall accuracy of 82.78%, demonstrating its effectiveness in entity grounding.

We notice that conventional object detection-based achieve moderate performance, with overall accuracies of 62.33% and 63.51%, respectively. Their performance is constrained by the limitations of object detection methods. Improvements can be observed when a more advanced object detection model is used (e.g., VinVL vs. RCNN). Knowledge-enhanced approaches have significantly improved performance. However, the absence of visual information in the incorporated knowledge imposes limitations on achieving further advancements.

The Closed-Source Advanced MLLMs, GPT-4o and GPT-4o-mini struggle significantly with person (PER) entities and have the lowest overall accuracies of 58.51% and 52.62% respectively. Their lack of task-specific fine-tuning likely impact their performance. In contrast, open-source MLLMs fine-tuned for this task generally outperform both object detection-based methods and closed-source models. Among them, Qwen2-VL-7B performs notably well with an overall accuracy of 80.77%. We further improve the performance by using ISR to train the MLLM to generate referring expressions, which bridge the gap between entity names and their visual regions. Compared to directly using Qwen2-VL-7B for predictions, our method boosts overall accuracy by 2.01%, with a notable 4.53% improvement in the MISC category.

4.3 Application to GMNER task

The entity grounding task is a crucial subtask of the recent proposed Grounded Multimodal Named Entity Recognition (GMNER, Yu et al., 2023) task, which aims to identify, classify and ground the entities in the text-image social post. The task formulation of GMNER is shown in Appendix D.To demonstrate the effectiveness and broad applicability of our method, we apply it to solve the GMNER task in conjunction with Multimodal Named Entity Recognition (MNER, Zhang et al., 2018) methods, which identifies and classifies named entities in the text by leveraging both textual and visual information. Here we employ two MNER methods, UMT (Yu et al., 2020) and PGIM (Li et al., 2023a), and compare the performance on the Twitter-GMNER dataset with previous methods.

The experiment results in Table 2 demonstrate the superiority of our EG method over the EG modules in previous GMNER methods. When using the same MNER methods, our approach achieves superior performance. For instance, UMT + Qwen2-VL-7B (ISR) + OFA$_{\text{Large}}$ improves the F1 score by 14.23% compared to UMT-VinVL-EVG, while PGIM + Qwen2-VL-7B (ISR) + OFA$_{\text{Large}}$ surpasses RiVEG (PGIM + OFA) by 5.29%. Even with a less advanced MNER model (i.e., UMT), the enhanced entity grounding performance enables us to reach near-state-of-the-art levels on the GMNER task. With a more advanced MNER model (i.e.,

Methods	w/ FT?	PER	LOC	ORG	MISC	Overall
Object Detection-Based						
RCNN-EVG (Yu et al., 2023)	✓	61.23	73.76	58.46	59.95	62.33
VinVL-EVG (Yu et al., 2023)	✓	62.59	77.48	59.25	58.69	63.51
Knowledge-Enhanced						
RiVEG (Li et al., 2024b)	✓	74.73	91.58	73.82	78.84	77.82
RiVEG (Augmented) (Li et al., 2024b)	✓	72.46	92.08	71.32	77.83	76.13
Closed-Source Advanced MLLMs						
GPT-4o	✗	35.51	89.36	71.32	70.53	58.51
GPT-4o-mini	✗	33.06	79.21	62.70	63.73	52.62
Fine-tuned Open-Source MLLMs						
InternVL2-8B (Chen et al., 2024)	✓	72.37	91.09	68.96	77.58	75.30
LLaVA-NeXT-7B (Liu et al., 2024)	✓	78.98	90.59	72.57	78.84	79.20
Qwen2-VL-7B (Wang et al., 2024b)	✓	80.52	91.83	**77.27**	75.82	80.77
Qwen2-VL-7B(Full SFT) + OFA$_{\text{Large}}$	✓	80.62	**92.82**	**77.27**	79.60	81.56
Qwen2-VL-7B (ISR) + OFA$_{\text{Large}}$	✓	**83.51**	92.33	76.96	**80.35**	**82.78**

Table 1: The performance comparison (Acc@0.5) of different methods on the dataset. Except for the closed-source advanced MLLMs, all other methods have been fine-tuned on the training set. Our proposed method demonstrates superiority over other approaches, achieving the highest accuracy for PER, LOC, MISC categories as well as in overall performance.

Methods	Twitter-GMNER		
	Pre.	Rec.	F1
Text			
HBiLSTM-CRF-None (Lu et al., 2018)	43.56	40.69	42.07
BERT-None (Devlin et al., 2019)	42.18	43.76	42.96
BERT-CRF-None	42.73	44.88	43.78
BARTNER-None (Yan et al., 2021)	44.61	45.04	44.82
Text+Image			
GVATT-RCNN-EVG (Lu et al., 2018)	49.36	47.80	48.57
UMT-RCNN-EVG (Yu et al., 2020)	49.16	51.48	50.29
UMT-VinVL-EVG (Yu et al., 2020)	50.15	52.52	51.31
UMGF-VinVL-EVG (Zhang et al., 2021a)	51.62	51.72	51.67
ITA-VinVL-EVG (Wang et al., 2022b)	52.37	50.77	51.56
BARTMNER-VinVL-EVG (Yu et al., 2023)	52.47	52.43	52.45
H-Index (Yu et al., 2023)	56.16	56.67	56.41
TIGER (Li et al., 2024c)	55.84	57.45	56.63
MQSPN (Tang et al., 2024)	59.03	58.49	58.76
RiVEG (PGIM + OFA) (Li et al., 2024b)	67.02	67.10	67.06
SCANNER (Ok et al., 2024)	68.34	68.71	68.52
UMT + Qwen2-VL-7B(Full SFT) + OFA$_{\text{Large}}$	64.26	63.43	63.84
PGIM + Qwen2-VL-7B(Full SFT) + OFA$_{\text{Large}}$	72.03	71.10	71.56
UMT + Qwen2-VL-7B (ISR) + OFA$_{\text{Large}}$	65.98	65.12	65.54
PGIM + Qwen2-VL-7B (ISR) + OFA$_{\text{Large}}$	**72.46**	**72.24**	**72.35**

Table 2: Experiment results on the Twitter-GMNER dataset.

PGIM), our approach outperforms the previous SOTA SCANNER model, with a 3.83% improvement in F1 score.

5 Analysis

5.1 Ablation Studies

Comparison with Supervised Fine-tuning In Table 3, we show the performance comparison of ISR with SFT training. We can observe that, with SFT initialization on a small amount of expert REs (Section 3.4.1), the model's performance is significantly enhanced. However, further SFT training on the remaining data yields only marginal gains. Starting from the same SFT-initialized model, ISR surpasses the performance of full-data SFT after just one iteration of training, with subsequent iterations leading to continued improvement in model performance until iteration 5.

Methods	PER	LOC	ORG	MISC	Overall
Qwen2-VL-7B	61.96	75.25	60.66	68.01	64.69
+ SFT (Initialization)	80.25	91.34	76.33	79.09	80.85
+ SFT (Full)	80.62	**92.82**	**77.27**	79.60	81.56
+ ISR (Iteration=1)	82.25	91.83	76.64	78.84	81.83
+ ISR (Iteration=2)	82.97	92.08	76.64	79.85	82.34
+ ISR (Iteration=3)	83.24	92.08	76.96	79.85	82.54
+ ISR (Iteration=4)	**83.51**	92.33	76.96	**80.35**	**82.78**
+ ISR (Iteration=5)	**83.51**	92.33	76.80	79.85	82.66

Table 3: The performance comparison of ISR at different iterations with SFT training.

Strategies of Preference Dataset Construction In our ISR training method, the strategy of preference dataset construction is essential for guiding the model's RE generation. In Figure 3, we compare our strategy stated in Section 3.4.2 with two variations: one without learning from successes and one without expert guidance. The former omits the mechanism that allows the model to learn from previously successful REs by always choosing ex-

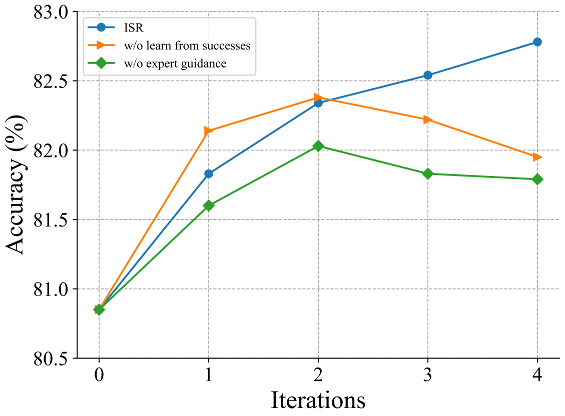

Figure 3: The performance comparison of different strategies of preference dataset construction. Iteration=0 represents the MLLM after SFT initialization.

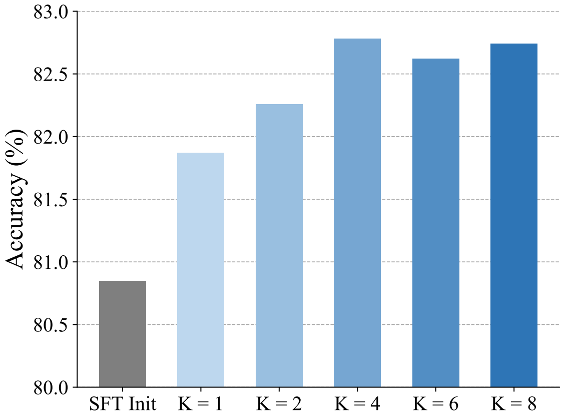

Figure 4: The model performance with different sampling number K.

pert REs as s_w. This limits the model's ability to reinforce successful exploration outcomes and results in over-fitting, as observed in the last few iterations. On the other hand, the variant without expert guidance removes the use of expert REs as fallback preferred responses, relying solely on self-generated REs. This strategy restricts the model's access to high-quality guidance when all generated REs are unsatisfactory,

5.2 Impact of the Sampling Number

We conduct experiments to explore the impact of the sampling number K, which represents the number of REs generated per entity prompt during each training iteration. As shown in Figure 4, the accuracy generally improves as K increases, with noticeable gains from $K = 1$ to $K = 4$. This indicates that too small K restricts the model's exploration, limiting its opportunities to generate diverse REs for each entity, which hampers the effective construction of preference data pairs. However, the accuracy gains start to plateau after $K = 4$, and increasing K will lead to additional computational burden and time costs. Based on our findings, $K = 4$ provides a balance, achieving high accuracy without excessive computational demands.

MLLM	VG	PER	LOC	ORG	MISC	Overall
Qwen2-VL-7B	OFA$_{\text{Large}}$	83.51	92.33	76.96	80.35	82.78
Qwen2-VL-2B	OFA$_{\text{Large}}$	81.16	92.33	76.33	80.10	81.56
LLaVA-NeXT-7B	OFA$_{\text{Large}}$	83.06	92.08	76.80	78.59	82.22
Llama-3.2-11B-Vision	OFA$_{\text{Large}}$	84.69	93.32	75.86	79.60	83.05
Qwen2-VL-7B	OFA$_{\text{Base}}$	83.24	92.57	77.12	79.09	82.54
Qwen2-VL-7B	Qwen2-VL-7B	84.24	92.57	77.27	81.11	83.33

Table 4: The performannce comparison of different MLLM and VG model selections.

5.3 Different MLLMs and VG Models

To further demonstrate the effectiveness and robustness of our approach, we conducted experiments with various MLLM and VG models, modifying only one component at a time while keeping the other at its default setting. The MLLM models include Qwen2-VL-2B, LLaVA-NeXT-7B and Llama-3.2-11B-Vision, and the VG models include OFA$_{\text{Base}}$ and Qwen2-VL-7B. As shown in Table 4, replacing the MLLM with similarly sized models results in stable performance. Even with a much smaller model, Qwen2-VL-2B, only a slight performance decline is observed, illustrating the robustness of our ISR training approach and its adaptability to various MLLM models. Meanwhile, replacing the MLLM with a larger model leads to improved performance. A similar trend is observed when changing the VG model: using the smaller OFA$_{\text{Base}}$ model leads to only minor performance drops, while switching to a more advanced VG model, Qwen2-VL-7B, yields performance gains.

6 Conclusion

In this paper, we present ISR, a novel training scheme to enhance the capability of multimodal large language models (MLLMs) in generating high-quality referring expressions (REs) for the entity grounding task. Through an iterative training process, the MLLM is encouraged to explore diverse ways of generating REs while learning from both successes and failures, without over-reliance on a fixed set of expert annotations. Extensive experiments validate the effectiveness and robustness of our method, achieving state-of-the-art performance in entity grounding and demonstrating its adaptability to broader multimodal tasks, such as GMNER. These findings highlight the potential of our approach to advance entity grounding methodologies, paving the way for more robust applications in multimodal knowledge graph construction and other complex multimodal understanding tasks.

Limitations

Despite its promising results, our proposed method has certain limitations that warrant further investigation. First, the reliance on a visual grounding model for feedback may constrain the system's performance to the accuracy and robustness of the grounding model employed. Future work could explore the integration of feedback from visual grounding outcome, advanced AI models, and human-beings to achieve a more comprehensive reward signal. Second, the method assumes the availability of high-quality training data, including entities and ground truth bounding boxes, which may not be feasible in all domains. Additionally, during our training process, sampling is required to construct preference data, which incurs additional time costs. Furthermore, DPO training demands a more substantial GPU memory compared to conventional SFT training. Future work could investigate more resource-efficient training methodologies.

Acknowledgements

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Key R&D Program of China (No. 2022YFC3600400).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2020. Visualsem: a high-quality knowledge graph for vision and language. *arXiv preprint arXiv:2008.09150*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023a. 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.

Jinyuan Li, Han Li, Di Sun, Jiahao Wang, Wenkun Zhang, Zan Wang, and Gang Pan. 2024b. Llms as bridges: Reformulating grounded multimodal named entity recognition. *arXiv preprint arXiv:2402.09989*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. 2023c. Transformer-based visual grounding with cross-modality interaction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–19.

Ziyan Li, Jianfei Yu, Jia Yang, Wenya Wang, Li Yang, and Rui Xia. 2024c. Generative multimodal data augmentation for low-resource multimodal named entity recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7336–7345.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. 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.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv 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.

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

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.

Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González, and Roberto J López-Sastre. 2017. Answering visual-relational queries in web-extracted knowledge graphs. *arXiv preprint arXiv:1709.02314*.

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

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, 25:2446–2458.

Jielong Tang, Zhenxing Wang, Ziyang Gong, Jianxing Yu, Shuang Wang, and Jian Yin. 2024. Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. *arXiv preprint arXiv:2407.21033*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

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

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

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

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

Fulong Ye, Yuxing Long, Fangxiang Feng, and Xiaojie Wang. 2023. Whether you can locate or not? interactive referring expression generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4697–4706.

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.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154, Toronto, Canada. Association for Computational Linguistics.

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

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

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

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

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. 2024. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735.

A Prompts Used in Our Work

A.1 Prompt for Expert REs Collection

We show the prompt for GPT-4o to generate the expert referring expression for a given entity. We first draw its ground-truth bounding box on the image, then prompt GPT-4o to generate an RE of the bounding box.

> Give a referring expression in one sentence, according to which the green bounding box in the image can be determined without any other prior knowledge.

A.2 Prompt for MLLM's RE Generation

We show the prompt for MLLM to generate REs.

> Given an image and a named entity, please check if the entity appears in the image. If it exists, please provide a description of the entity in the image.
> Image: {image}
> Entity: {entity (type)}

A.3 Prompt for MLLM's Direct Entity Grounding

We fine-tune several open-source MLLMs on the entity grounding training set as baselines and prompt them to direct solve the entity grounding task.

> Given an image and a named entity, please check if the entity appears in the image. If it exists, please provide the corresponding bounding box; otherwise, please answer "The entity does not appear in the image."
> Image: {image}
> Entity: {entity (type)}

B Detailed Statistics of the Dataset

Split	#Entity	#Groundable Entity	#Image
Train	11457	4449	6565
Dev	2406	964	1405
Test	2543	1046	1500
Total	16406	6459	9470

Table 5: The statistics of the entity grounding dataset.

We utilize the images, entities and their bounding box annotations in Twitter-GMNER dataset (Yu et al., 2023) as the entity grounding dataset for training and evaluation. The statistics of the dataset are shown in Table 5. It is noteworthy that we manually filter out entities with incorrect bounding box annotations in the training and development sets of the Twitter-GMNER dataset. To maintain the fairness and integrity of the evaluation, all entities in the test set were retained without modification.

C Training and Inference Costs

The detailed training and inference costs are shown below:

- GPU Hours: The training was conducted on 1 NVIDIA A800 80G GPU. SFT initialization stage took about half an hour. Each iteration of the ISR training process, including preference data construction and DPO training, took approximately 3 hours, and the total training time was about 12 GPU hours.
- Expert Annotation Cost: The expert REs were generated using GPT-4o, with a total cost of approximately \$20.
- Closed-source Model API Cost: We prompt GPT-4o and GPT-4o-mini to directly solve entity grounding task as our baselines. The total cost for GPT-4o was \$10 and the total cost for GPT-4o-mini was \$0.6.

D Task Formulation of GMNER

Given a sentence $S = \{s_1, ..., s_n\}$ and its corresponding image I, the goal of GMNER is to extract a set of multimodal entity triples:

$$Y = \{(e_1, c_1, v_1), ..., (e_m, c_m, v_m)\} \qquad (9)$$

where e_i is the i-th entity in the sentence, $c_i \in \{PER, LOC, ORG, MISC\}$ refers to the type of e_i, and v_i denotes the visually grounded region of e_i. If e_i is ungroundable, v_i should be $None$; otherwise, v_i consists of a 4D coordinates representing the top-left and bottom-right locations of the grounded bounding box, i.e., (x_1, y_1, x_2, y_2).

E Case Study

Here, we further provide a detailed comparison of the predictions given by knowledge-enhanced method RiVEG (Li et al., 2024b) and our proposed method for challenging samples, which illustrates how refined REs generated by the MLLM trained

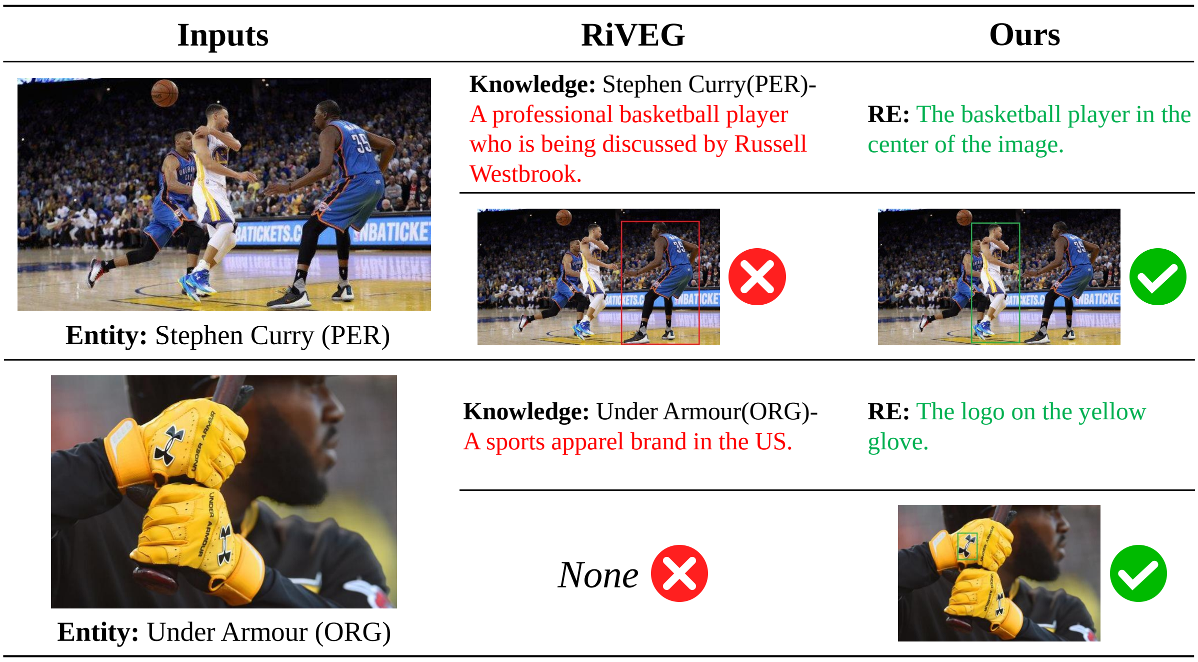
Inputs
RiVEG
Ours
Knowledge: Stephen Curry(PER)-A professional basketball player who is being discussed by Russell Westbrook.
RE: The basketball player in the center of the image.
Entity: Stephen Curry (PER)
Knowledge: Under Armour(ORG)-A sports apparel brand in the US.
RE: The logo on the yellow glove.
None
Entity: Under Armour (ORG)

Figure 5: A case study on the entity grounding predictions given by RiVEG and our method.

with ISR benefit the process of entity grounding, providing fine-grained visual information for the grounding model. The cases are shown in Figure 5.

In the first case, the entity to be grounded is “Stephen Curry (PER)” in a basketball game image. The RiVEG method, leveraging knowledge: “a professional basketball player who is being discussed by Russell Westbrook”, fails to correctly locate Stephen Curry in the image. This error arises because the factual description focuses on textual context and lacks explicit visual clues to unambiguously ground the entity. In contrast, our method generates the RE: “The basketball player in the center of the image”, which explicitly describes the spatial position of Stephen Curry in the image. This precise RE enables the visual grounding model to accurately identify the bounding box for Stephen Curry.

The second example involves grounding the entity “Under Armour (ORG)”, a sports apparel brand, in an image featuring gloves with the Under Armour logo. RiVEG relies on the factual description “a sports apparel brand in the US”, which provides no direct visual information. Consequently, the model fails to ground the entity. Our method generates the RE: “The logo on the yellow glove”, effectively linking the entity to a specific visual attribute (the logo) and its associated object (the glove). This RE facilitates the grounding model to locate the correct bounding box around the logo.