Granular Entity Mapper: Advancing Fine-grained Multimodal Named Entity Recognition and Grounding

Anonymous ACL submission

Abstract

Multimodal Named Entity Recognition and Grounding (MNERG) aims to extract paired textual and visual entities from texts and images. It has been well explored through a two-step paradigm: initially identifying potential visual entities using object detection methods and then aligning the extracted textual entities with their corresponding visual entities. However, when it comes to fine-grained MNERG, the long-tailed distribution of textual entity categories and the performance of object detectors limit the effectiveness of traditional methods. Specifically, more detailed classification leads to many low-frequency categories, and existing object detection methods often fail to pinpoint subtle regions within images. To address these challenges, we propose the **G**ranular **E**ntity **M**apper (GEM) framework. Firstly, we design a multi-granularity entity recognition module, followed by a reranking module based on the Multimodal Large Language Model (MLLM) to incorporate hierarchical information of entity categories, visual cues, and external textual resources collectively for accurate fine-grained textual entity recognition. Then, we utilize a pre-trained Large Visual Language Model (LVLM) as an implicit visual entity grounder that directly deduces relevant visual entity regions from the entire image without the need for bounding box training. Experimental results on the GMNER and FMNERG datasets demonstrate that our GEM framework achieves state-of-the-art results on the fine-grained content extraction task.

1 Introduction

Multimodal Named Entity Recognition and Grounding (MNERG) aims to recognize named entities and corresponding image regions from multimodal data, which is crucial for various applications, including multimodal knowledge graph construction, video recommendation, and multimodal chatbot. Typical MNERG approaches often

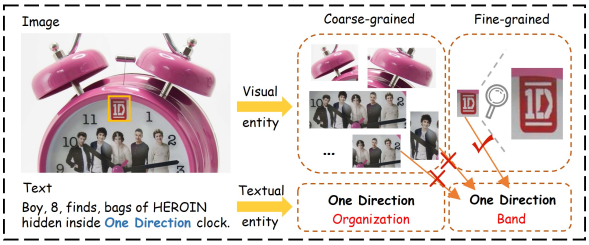

Figure 1: An example to illustrate the fine-grained MNERG. The textual entity is annotated by highlighting, and the visual entity is annotated by the bounding box.

involve a two-step framework (Yu et al., 2023), where a well-trained object detection model is utilized to extract image regions as potential visual entities. Then, a cross-modality modeling framework is leveraged to extract and link textual entities with corresponding potential visual entities, enabling multimodal entity alignment. Along this line, numerous efforts have been recently dedicated to exploring this problem, and notable performances have been achieved.

Moreover, to better capture the complexity of the real world, fine-grained MNERG endeavors to classify textual entities into more detailed categories and extract smaller, more precise visual entity regions. Indeed, delving into fine-grained MNERG reveals new challenges and limitations. On the one hand, **fine-grained textual entities** often suffer from the problem of long-tailed distribution, necessitating external information sources to achieve precise recognition and classification of these textual entities. On the other hand, **fine-grained visual entities** often exhibit a wide variety of sizes, which challenges traditional object detection methods in consistently recalling them and further hinders multimodal entity alignment. For example, as shown in Figure 1, the textual entity *One Direction* requires common knowledge about the band and the individuals in the image to help discriminate it from other organization categories. Additionally, existing object detection methods can only detect coarse-grained potential visual entity regions in the figure, and the logo as the corresponding visual entity does not appear among the candidates due to its small size. Therefore, supplementing the valuable knowledge and clues and tracing relevant regions directly from the images is essential for fine-grained content extraction.

Fortunately, recent years have witnessed the prosperity of multimodal large models (Li et al., 2022, 2023b; Liu et al., 2023), which have shown advanced capabilities in comprehending relationships and reasoning in complex scenarios involving texts and images. Inspired by such progress, we fully utilize the cross-modal interacting capabilities of various multimodal large models and propose a novel fine-grained MNERG framework, named Granular Entity Mapper (GEM), to address the above challenges.

Firstly, we employ a knowledge-enhanced multi-granular entity recognition module, followed by a multimodal reranking module, to incorporate external textual knowledge, structured information, and visual cues collectively for accurate fine-grained textual entity recognition. Specifically, we acquire rich external knowledge from Large Language Models (LLMs) through prompts and then preliminarily recognize entities constrained by the entity category hierarchy to enhance long-tailed categories. Leveraging the powerful relationship comprehension and endogenous multimodal knowledge of Multimodal Large Language Models (MLLMs[1]), we rerank the predicted textual entity categories to differentiate long-tailed categories from similar categories. Secondly, we utilize a Large Visual Language Model (LVLM) as an implicit grounder to establish associations between textual entities and their corresponding visual entity regions, enabling the recognition of visual entities even without training on annotated bounding boxes. Due to the numerous natural text and image alignments during the pre-training stage, our grounder is suitable for open-vocabulary textual entities and can directly identify the corresponding regions across the image, overcoming the limitations associated with traditional object detectors for fine-grained visual entity grounding.

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

- We propose leveraging multi-granularity, multi-perspective information to enhance the recognition of fine-grained textual entities.
- We propose employing an implicit paradigm to effectively pinpoint fine-grained visual entity regions directly from images, eliminating the reliance on preliminary object detection.
- Extensive experiments show that our framework achieves state-of-the-art results on the GMNER and FMNERG datasets and significantly improves fine-grained entity extraction.

2 Related Work

2.1 Multimodal Named Entity Recognition

Multimodal Named Entity Recognition is a pivotal task designed to extract entities from social media texts with the help of images. Previous approaches in MNER could be broadly categorized into two types: (1) Modal-Interaction based: BMA (Moon et al., 2018) and ADACAN (Zhang et al., 2018) utilized various attention mechanisms to establish relationships between texts and images. UMT (Yu et al., 2020) pioneered using a multimodal transformer for this task, while CAT (Wang et al., 2022c) further refined cross-attention representation by incorporating label semantics. (2) Knowledge-based: ITA (Wang et al., 2022b) extracted sample knowledge from images and MoRe (Wang et al., 2022a) went a step further by retrieving information from Wikipedia. PGIM (Li et al., 2023a) had stood out by using demonstrations to extract implicit knowledge from LLMs.

2.2 Entity Grounding

Entity grounding involves ascertaining the relevance of a textual entity to an image and pinpointing the most probable region where it appears. Previous methods (Wang et al., 2023; Yu et al., 2023) used a Cross-Modality Transformer (CMT) to calculate the similarity between extracted textual entities and candidate visual entities identified by object detection (Zhang et al., 2021b; Girshick, 2015). H-index and Tiger (Wang et al., 2023; Yu et al., 2023) used a special token to represent the relationships between textual entities and images, facilitating the matching of candidate visual entities.

2.3 Multimodal Named Entity Recognition and Grounding

This task integrates multimodal named entity recognition with entity grounding to extract structured

[1]In this paper, MLLM refers to the training of multimodal large models aligned with large language models, whereas LVLM primarily undergoes typical multimodal pre-training.

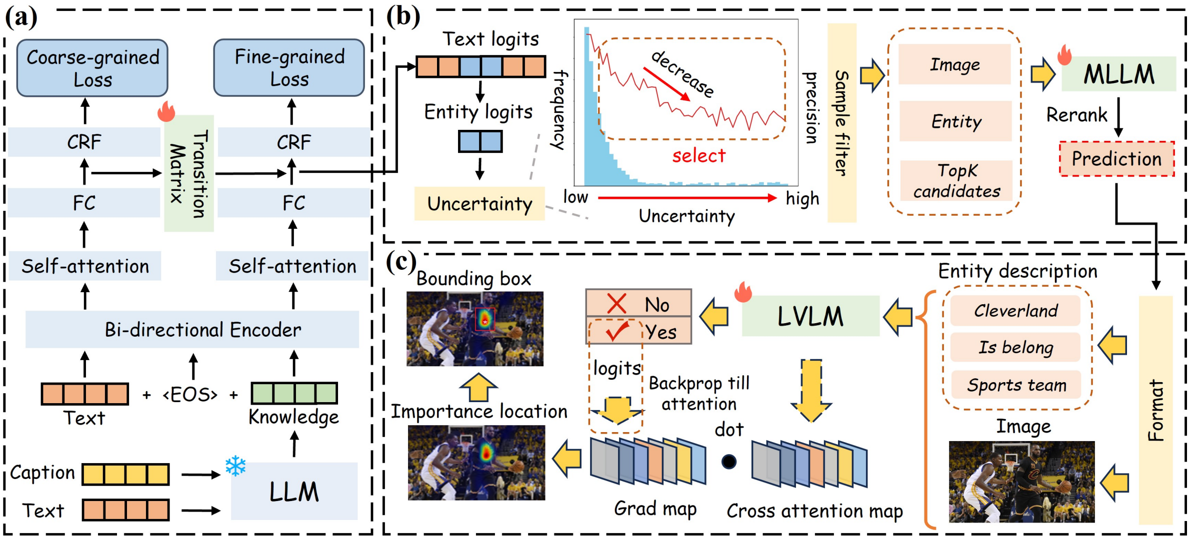

Figure 2: The overall framework of our GEM. (a) Knowledge-enhanced multi-granularity textual entity recognition. (b) MLLM-based textual entity category reranking. (c) LVLM-based implicit visual entity grounding.

information from texts and images simultaneously. It combines the above-mentioned methods and follows a two-step paradigm. Additionally, H-index and Tiger (Wang et al., 2023; Yu et al., 2023) introduced a new paradigm that used a special token to predict the relevance between textual and visual entities. Among them, Tiger achieved certain improvements in fine-grained textual entity recognition by simultaneously predicting labels at both coarse and fine granularities. However, along with previous methods, they grappled with the long-tailed distribution of fine-grained categories and lacked valid candidate regions for fine-grained visual entities. Meanwhile, previous knowledge-based methods (Wang et al., 2022a; Li et al., 2023a) either introduced misleading noise or required numerous manually annotated samples, making it difficult to aid fine-grained textual entity recognition. Our work integrates multi-granularity, multi-perspective information to deeply mine fine-grained textual entities and directly extracts the visual region from the image rather than relying on predefined candidates.

3 Method

In this section, we first formulate the fine-grained MNERG task and then explain our framework in detail. Our GEM comprises three main modules: (1) The *Knowledge-enhanced multi-granularity textual entity recognition module* first leverages external auxiliary knowledge and the hierarchical structure of entity categories to preliminarily recognize textual entities. (2) The *MLLM-based textual entity category reranking module* comprehensively utilizes multimodal clues extracted by cross-modality interaction for accurate entity category prediction, combined with a filtering regime. (3) The *LVLM-based implicit visual entity grounding module* utilizes an LVLM to match textual and visual entities.

3.1 Problem Formulation

Given a sentence T and the associated image I, the goal of fine-grained MNERG is to extract a set of triples S expressed as:

$$\{(e_1, c_1, o_1), (e_2, c_2, o_2), \ldots, (e_N, c_N, o_N)\}, \quad (1)$$

where e_i represents the i-th textual entity in sentence T, c_i represents the category of textual entity e_i, o_i represents the visual entity region corresponding to textual entity e_i in image I, N represents the number of textual entities in sentence T. If the textual entity has a corresponding visual entity in the image, o_i is a four-dimensional vector containing the coordinates of the bounding box; otherwise, o_i is $None$. o_i can be expressed as:

$$o_i = \begin{cases} \text{None}, & \text{ungrounded}, \\ (x_1^i, y_1^i, x_2^i, y_2^i), & \text{grounded}, \end{cases} \quad (2)$$

where (x_1^i, y_1^i) and (x_2^i, y_2^i) separately represent the top-left and bottom-right coordinates of the bounding box for the i-th entity.

3.2 Knowledge-enhanced Multi-granularity Textual Entity Recognition Module

To augment the long-tailed textual entity category with valuable knowledge, we employ an LLM to incorporate external auxiliary knowledge. Subsequently, we utilize a modified multi-granularity NER model to recognize textual entities by integrating the entity category hierarchy.

3.2.1 Knowledge Augmentation

With the help of the LLM's internal knowledge, valuable information is provided to support both entity classification and span recognition, thereby enhancing the model's ability to identify out-of-vocabulary textual entities such as *Redmi R7*. Specifically, we concatenate the text with the corresponding image caption acquired by BLIP-2 (Li et al., 2023b) and feed them into the LLM with designed $Instruction$ to obtain the auxiliary knowledge. Subsequently, we concatenate the text with the acquired knowledge using a special token $\langle SEP\rangle$ to delineate them, as expressed:

$$(t_1, t_2, \ldots, t_{N_1}, \langle SEP\rangle, a_1, a_2, \ldots, a_{N_2}), \quad (3)$$

where t_i represents the input token of text, a_i is the auxiliary knowledge token, which is then fed into a modified NER model for encoding and getting the representation of the sequence:

$$(y_1, y_2, \ldots, y_{N_1}, y_{N_1+1}, \ldots, y_{N_1+N_2+1}). \quad (4)$$

3.2.2 Multi-Granularity Prediction

As shown in Figure 2 (a), we have modified the typical NER model into a dual-path structure with independent parameters, enabling simultaneous predictions at both coarse and fine granularity. Specifically, we set different output dimensions of the fully connected layer to map various granularities, while a Conditional Random Field (CRF) (Huang et al., 2015) layer refines the sequence labeling. We define the probability of the label sequence c given the input sentence T, so the CRF refine the labels can be expressed as:

$$P(c|T) = \frac{\prod_{i=1}^{N_1+N_2+1} \psi(c_{i-1}, c_i, y_i)}{\sum_{c' \in C} \prod_{i=1}^{N_1+N_2+1} \psi(c'_{i-1}, c'_i, y_i)}, \quad (5)$$

where $\psi(c_{i-1}, c_i, y_i)$ and $\psi(c'_{i-1}, c'_i, y_i)$ are potential functions. We use the negative log-likelihood as the loss function for the input sequence with gold labels c^* for different granularities:

$$L^c_{NLL}(\theta) = -\log P_\theta(c^*_c|S), \quad (6)$$

$$L^f_{NLL}(\theta) = -\log P_\theta(c^*_f|S), \quad (7)$$

$$L_{NLL} = \alpha L^c_{NLL} + (1-\alpha) L^f_{NLL}, \quad (8)$$

where L^c_{NLL} and L^f_{NLL} respectively represent the loss for coarse and fine granularity and α is the weight coefficient to balance the losses.

3.2.3 Multi-Granularity Augmentation

We will now describe how multi-granularity information improves predictions for long-tailed categories. The logit prediction within the coarse-grained categories is extracted, and a learnable transition matrix is utilized to boost the probabilities of corresponding fine-grained categories. Specifically, we denote the logit prediction by the fully connected layer within the coarse-grained categories as $(y^c_1, y^c_2, \ldots, y^c_{N_1+N_2+1})$ and fine-grained logit prediction as $(y^f_1, y^f_2, \ldots, y^f_{N_1+N_2+1})$, where $y^c_i \in \mathbb{R}^{C_c}$ and $y^f_i \in \mathbb{R}^{C_f}$. Here, C_c and C_f represent the number of coarse and fine granularity categories, respectively. Then, a learnable transition matrix $M \in \mathbb{R}^{C_c \times C_f}$ transitions y^c_i and adds it to y^f_i with a weight β:

$$y^f_i = \beta M y^c_i + (1-\beta) y^f_i. \quad (9)$$

Notably, M is initialized with the co-occurrence frequency of coarse and fine granularity categories and then normalized.

3.3 MLLM-based Textual Entity Category Reranking Module

For further differentiation of long-tailed categories from others based on previous granularity augmentation, we employ the MLLM as a multimodal reranker combined with a sample filtering mechanism to refine appropriate samples.

3.3.1 Sample Filter and Selection

Previous findings (Zhang et al., 2024; Ma et al., 2023) have revealed that LLMs are suitable for hard samples. Inspired by them, we filter and select such challenging samples for further processing. Specifically, we extract textual entity embeddings $(y_{e^1_i}, y_{e^2_i}, \ldots, y_{e^M_i})$ and pool these tokens to form the textual entity's representation. Here, e^j_i represents the j-th token of the i-th textual entity. We then merge the logits of B-I within the same category and apply softmax to represent the probabilities $(p(x^{e_i}_1), p(x^{e_i}_2), \ldots, p(x^{e_i}_{C_f}))$ of each category. Subsequently, we calculate the information entropy $H(p)$ of the distribution to evaluate the difficulty associated with the textual entity as follows:

$$H(p(e_i)) = -\sum_j p(x^{e_i}_j) \log p(x^{e_i}_j). \quad (10)$$

Using a predefined threshold γ, we filter and further process samples with information entropy that exceeds this value. Notably, we consider the remaining samples to be well-processed by the previous modules and not require further processing.

3.3.2 Entity Category Reranking

To avoid excessive textual entity categories from interfering with the MLLM, we select the $topK$ categories with the highest probabilities as candidates, based on the predicted probabilities $(p(x_1), p(x_2), \ldots, p(x_{C_f}))$. The sample is then formatted as $(Instruction, I, T, candidates)$ and input into the instruction-tuned MLLM for reranking to select the best category. Actually, the candidates usually belong to the same coarse-grained categories due to the multi-granularity augmentation. Therefore, the long-tailed categories can be further differentiated from similar categories.

To instruction-tune the MLLM, we construct a candidate set of length K including the golden label, $K-2$ fine-grained categories within the same coarse-grained category, and one distinct category from a different category. This enhances robustness by accounting for occasional misclassifications of the coarse-grained category by the model.

3.4 LVLM-based Implicit Visual Entity Grounding Module

Visual entity grounding involves two primary steps: confirming the relevance of a textual entity to an image and precisely grounding the visual region within the image. Consequently, an LVLM is trained on the relevance between entities and images and subsequently infers the grounding regions using an implicit paradigm. Notably, to align with the labeling method of visual entities, we generate bounding boxes for grounding positions using a visual prompt model.

3.4.1 Textual Entity-Image Matching

We finetune an off-the-shelf LVLM (BLIP) (Li et al., 2022) equipped with its Image-Text-Match head serving as a binary classifier to determine the textual entity's relevance (P_T, P_F) to the image. Here, P_T denotes the probability that the entity matches the picture, and P_F denotes the probability that it does not. Meanwhile, we construct a dataset formulated as $(e_i, Instruction, c_i, I, label)$ to finetune our model. The label is a boolean value indicating whether the corresponding visual entity is present in the image. We include entity categories because entities sharing the same name but belonging to different categories may represent different elements in the image, such as the athlete *Jordan* and the brand *Jordan*.

3.4.2 Visual Entity Tracing

In fact, we can trace the visual entity's position to explain why the classifier identifies the textual entity relevant to the image. For the textual entity determined to be relevant to the image, we extract P_T and apply gradient-based weighting (Selvaraju et al., 2017; Tiong et al., 2022) to the cross-attention maps, deriving importance scores for various regions within the image as follows:

$$s_i = \frac{1}{H} \sum_{j=1}^{S} \sum_{h=1}^{H} \max(0, \frac{\partial P_T}{\partial A_{ji}^{(h)}}) A_{ji}^{(h)}. \quad (11)$$

Here, H refers to the total count of attention heads, S denotes the overall length of the tokens, and $A_{ji}^{(h)}$ denotes the attention score between the i-th patch and the j-th token within the h-th attention head. We then resize the score map to match the size of the original image, allowing us to assess the importance of each region. Having obtained the importance distribution of the image regions associated with the textual entity, we consider the region with the highest importance score as the potential key visual entity linked to the textual entity. This process effectively establishes a connection between the textual entity and the relevant visual region within the image.

3.4.3 Bounding Box Generation

Visual entities are typically represented using bounding boxes. Therefore, we need to transform the importance distribution into specific coordinates. However, there is often a discrepancy between the identified importance region and the target bounding box. We must deduce the bounding box from the region of local importance. SEEM (Zou et al., 2023) is a visual prompt model that can separate the object using a pointed hint to generate its mask. Therefore, we use it to isolate the entity object based on the coordinates of the highest score point within the score map. Subsequently, we derive the bounding box coordinates as our final prediction based on the generated mask. During this process, we deduced the grounding region of the visual entity solely based on the relationships between textual entities and images, thus eliminating the need for training with extensive hand-annotated bounding boxes in the dataset.

Modality	Methods	GMNER			FMNERG		
		MNERG	MNER	EEG	MNERG	MNER	EEG
Text	HBiLSTM-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
	BART / T5-Paraphrase-None	44.82	79.83	48.99	37.33	65.07	48.97
Text+Image	GVATT-OD-EVG	48.57	76.26	53.32	40.32	60.35	54.35
	UMT-OD-EVG	50.29	78.58	54.78	41.32	61.63	54.43
	UMGF-OD-EVG	51.67	78.83	55.74	41.92	61.79	54.75
	ITA-OD-EVG	51.56	79.37	55.69	42.78	63.21	57.26
	BART / MMT5-OD-EVG	52.45	80.39	55.66	45.21	66.61	58.18
	H-Index / TIGER	56.41	79.73	61.18	46.55	64.91	61.96
	GEM (BERT)	59.83 ± 0.21	83.15 ± 0.12	63.16 ± 0.09	50.54 ± 0.19	68.09 ± 0.15	63.59 ± 0.07
	GEM (RoBERTa)	**61.54 ± 0.17**	**84.81 ± 0.06**	**64.49 ± 0.10**	**52.48 ± 0.14**	**70.80 ± 0.11**	**65.52 ± 0.05**

Table 1: Performance comparison between GEM and all the baselines. Results for all baselines are sourced from Wang et al. (2023); Yu et al. (2023), and the best results are highlighted in bold. Importantly, we utilize VinVL (Zhang et al., 2021b) as the main object detection method, denoted as OD, and employ RCNN (Girshick, 2015) in some baseline evaluations of the GMNER dataset. The mean and standard deviation across all the metrics are obtained through three random runs.

4 Experiments

4.1 Settings

Datasets We conducted experiments using two public MNERG datasets: GMNER and FMNERG. Notably, the GMNER dataset includes only four coarse-grained categories for textual entities, whereas the FMNERG dataset labels eight coarse-grained and fifty-one fine-grained categories. More details are in Appendix A.

Baselines To evaluate the performance of our framework in FMNERG, we benchmarked our approach with the following baselines: (1) Text-only: (Huang et al., 2015; Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020) Only extracting textual entities. (2) EVG-based: (Jia et al., 2023; Yu et al., 2020; Wang et al., 2022b) Extracting textual entities, then selecting corresponding visual entities. (3) Unified-Generative: (Wang et al., 2023; Yu et al., 2023) Simultaneously capturing textual and corresponding visual entities with a multi-modality generative model. More details are in Appendix B.

Evaluation Referring to prior work, we assessed our framework's performance across three distinct subtasks. (1) Multimodal Named Entity Recognition (MNER) involves predicting the correct textual entity spans and their types. (2) Entity Extraction & Grounding (EEG) entails identifying both the textual entity spans and their corresponding visual entities. We apply a threshold of 0.5 for filtering Intersection over Union (IoU) scores between ground truth and predicted bounding boxes. (3) Multimodal Named Entity Recognition and Grounding (MNERG) comprehensively evaluate the performance of both MNER and EEG, ensuring the accuracy of the triplet (e_i, c_i, o_i). All subtasks were evaluated using the F1-score.

Implementations All model components run on a single NVIDIA RTX 4090 GPU using PyTorch. We set $\alpha = 0.1, \beta = 0.1$ for textual entity recognition and selected ChatGPT as our knowledge base. Additionally, we set $\gamma = 0.2$ for sample filtering and employed LoRA with rank $= 64$ to instruction-tune LLaVA (Liu et al., 2023) for reranking. The BLIP (Li et al., 2022) was fine-tuned to assess the relevance between textual entities and images. To ensure fair comparisons, we present results using both BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as backbone networks. Since the GMNER dataset contains only coarse-grained textual entity categories, we removed the multi-granularity module and ensured that all categories were considered during reranking. More details are in Appendix C.

4.2 Comparison with Baselines

The performance comparison of our GEM and the baselines is detailed in Table 1. We have the following observations: (1) Our GEM consistently achieves the best performance across all subtasks using both BERT and RoBERTa, with a maximum absolute improvement of 5.13% and 5.93% for the entire assessment in the GMNER and FMNERG datasets, respectively. This indicates that our model provides additional capabilities beyond those of the backbone models. (2) In multimodal named entity

Methods	Coarse-grained			Fine-grained		
	Pre	Rec	F1	Pre	Rec	F1
Textual entity						
Base model	80.92	82.89	81.89	66.79	67.40	67.10
Multi	**81.37**	83.29	82.32	67.74	68.56	68.15
Rerank	81.07	82.99	82.02	68.92	69.64	69.28
Multi+Rerank	81.23	**83.49**	**82.34**	**70.25**	**71.36**	**70.80**
Visual entity						
CMT-RCNN	63.89	62.94	63.41	16.70	15.35	16.00
CMT-VinVL	63.47	62.02	62.73	18.71	17.08	17.86
GEM-wo	62.39	63.10	62.74	25.77	26.25	26.01
GEM	**66.29**	**67.04**	**66.66**	**35.64**	**36.38**	**36.01**

Table 2: Performance comparison across different granularities in textual entity recognition and visual entity grounding. Evaluations are based on precision, recall, and F1-score. The term "Multi" denotes the module that incorporates multi-granularity information.

recognition, our model achieves a 4.19% higher score than the previous best result in the FMNERG dataset, demonstrating its ability to capture textual entities at a finer granularity level. (3) In entity extraction and grounding, we achieve obvious improvements that surpass the progress in entity span predictions across all datasets. This proves that even without training with bounding boxes, we can accurately identify visual entities and link them to corresponding textual entities.

4.3 Fine-grained Content Performance

We compared textual entity recognition and visual entity grounding across various modules and granularities within the FMNERG dataset to validate our approach's effectiveness on fine-grained content.

In fine-grained textual entity recognition, we employed a typical NER model with auxiliary knowledge as the base model. Then we evaluated the effects of refining the base model's results either by incorporating multi-granularity information or by using a reranking module. As shown in table 2, fine-grained categories exhibit more remarkable improvement compared to coarse-grained categories, demonstrating that the performance enhancement in fine-grained categories stems from a better comprehension of detailed content across different modalities rather than a general enhancement. Multi-granularity information primarily boosts the logit prediction of long-tailed categories without directly distinguishing them from others. However, it provides better base candidates for reranking and further differentiates the long-tailed category from other similar categories. Combining them leads to cooperative improvement.

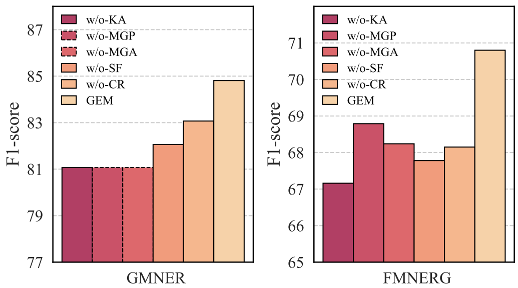

Figure 3: Performance comparison between GEM and its variants. We omit the MGP and MGA components and represent them with dashed lines aligned with AK values for consistent comparison in the GMNER dataset.

In fine-grained visual entity grounding, we formulated the visual entity with an area less than one-fiftieth of the image as the fine-grained visual entity. The Cross Modality Transformer (CMT) was selected as our base model, which effectively linked textual entities to their corresponding visual entities identified by object detection. Various object detection (Girshick, 2015; Zhang et al., 2021b) methods were employed to support CMT. Notably, the model variant GEM-wo represents our approach using the same initial model weights but without training under the textual entity-image matching task. From Table 2, it is evident that our GEM and its variant significantly outperform the typical method in fine-grained visual entity grounding by a large margin. This superior performance is due to the direct grounding of visual entities across the entire image with strong text-object alignment capability, breaking away from previous non-end-to-end grounding processes. Additionally, we note that our GEM performs better than its variant, indicating that our textual entity-image matching significantly enhances the alignment between textual and visual entities, rather than relying solely on the text-image alignment from the pre-training stage.

4.4 Ablation Analysis

To verify the effectiveness of each design in our model, we compared GEM with five variants evaluated on the MNER subtask:

- **w/o-KA** removes knowledge augmentation.
- **w/o-MGP** removes multi-granularity prediction.
- **w/o-MGA** removes multi-granularity augmentation (excluding the transition matrix).
- **w/o-SF** removes sample filter.
- **w/o-CR** removes category reranking.

According to the results shown in Figure 3, GEM outperforms all its variants. Specifically, the w/o-

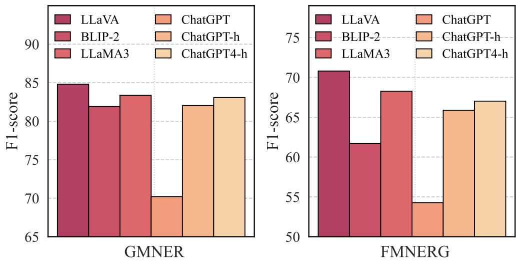

Figure 4: Performance comparison across different models in textual entity category reranking.

Methods	Coarse-grained			Fine-grained		
	Pre	Rec	F1	Pre	Rec	F1
LLaVA	54.88	55.64	55.26	21.59	22.01	21.80
BLIP-2	61.98	61.19	61.58	29.06	28.69	28.87
BLIP	**66.29**	**67.04**	**66.66**	**35.64**	**36.38**	**36.01**

Table 3: Performance comparison with LLaVA, BLIP-2, BLIP in visual entity grounding.

KA underperforms compared to other variants, highlighting that the base model's performance sets the upper limit for textual entity recognition. Since NER is a strict matching problem, providing the valuable knowledge not only enhances span prediction but also boosts the logit prediction for relevant entity categories. Meanwhile, we can see that w/o-MGA shows a relative performance degradation compared to w/o MGP, proving that fine-grained logit augmentation is essential for deriving extra knowledge from coarse-grained information. Besides, we observe a performance decrease when removing the sample filter, illustrating that the base and reranking models have different expertise in textual entity recognition. Therefore, combining them is crucial to enhance the final results. Notably, the performance degrades when we discard the reranking, indicating a necessity for the MLLM to provide essential multimodal knowledge to help distinguish the textual entity.

5 Discussion

In this section, we detail our preference for using the MLLM with instruction-tuning for reranking instead of a larger model with in-context learning. Furthermore, our results show that the BLIP outperforms existing MLLMs in visual entity grounding. More discussions are in Appendix D, E, F.

5.1 Different Models for Reranking

We compared the reranking capabilities across various modalities and sizes of models, feeding text-only models with captions instead of images. Specifically, we used in-context learning to prompt GPT models, and the "$-h$" notation indicates that we provided heuristic candidate logit predictions to the models to avoid overconfidence in their internal knowledge like prophet (Shao et al., 2023).

According to Figure 4, we can see that LLaVA performs best across all models, indicating that the acquisition of additional multimodal information aids in comprehending the meaning of samples. LLaMA3 outperforms BLIP-2 due to its superior instruction-following and text comprehension capabilities during the pre-training stage. However, the GPT series exhibits a remarkable decline in performance within the few-shot setting, even with heuristic hints. This demonstrates that in-context learning struggles to grasp the reranking paradigm for entity classification, highlighting the superiority of our instruction-tuning reranking paradigm.

5.2 Different Models for Visual Grounding

To illustrate why we chose BLIP as the implicit visual entity grounder, we instruction-tuned widely used MLLMs (LLaVA, BLIP-2) to assess the relevance between textual entities and images. Subsequently, we extracted P_T to weight the feature maps in the visual encoder appropriately.

As shown in Table 3, BLIP consistently outperforms other MLLMs across all scores. This superiority can be attributed to two main factors: (1) Alignment Bias. MLLMs typically align the visual embeddings with the text rather than with the original image, introducing biases in visual entity grounding. (2) Alignment Absence. MLLMs are mainly trained with generation loss to align with the text, which makes it difficult to extract effective region-specific information and tends to distribute the information across the entire image.

6 Conclusion

In this paper, we introduced GEM, a novel framework for fine-grained multimodal named entity recognition and grounding based on integrated multi-granularity and multi-level information. By harnessing the rich multimodal knowledge and linguistic understanding from multimodal pre-training, we enhanced the comprehension of fine-grained information in both images and texts. Extensive experimental results demonstrated the superior performance of the GEM framework.

7 Limitations

We briefly mention some limitations of our work. First, we have adopted caption information for preliminary entity recognition, however this may lead to missing information and introduce noise into the subsequent reranking process. Moreover, although our grounding paradigm demonstrates remarkable performance for fine-grained visual entities, it faces challenges when pinpointing certain very large regions, revealing a gap in our box generation method.

References

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ross B. Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.

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

Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. MNER-QG: an end-to-end MRC framework for multimodal named entity recognition with query grounding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8032–8040. AAAI Press.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jinyuan Li, Han Li, Zhuo Pan, and Gang Pan. 2023a. Prompt chatgpt in MNER: improved multimodal named entity recognition method based on auxiliary refining knowledge from chatgpt. *CoRR*, abs/2305.12212.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1990–1999. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10572–10601. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 852–860. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14974–14983. IEEE.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C. H. Hoi. 2022. Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 951–967. Association for Computational Linguistics.

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, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3934–3943. ACM.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022a. Named entity and relation extraction with multi-modal retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5925–5936. Association for Computational Linguistics.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022b. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3176–3189. Association for Computational Linguistics.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022c. CAT-MNER: multimodal named entity recognition with knowledge-refined cross-modal attention. In *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*, pages 1–6. IEEE.

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, ACL 2020, Online, July 5-10, 2020*, pages 3342–3352. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9141–9154. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14347–14355. AAAI Press.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE.

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, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5674–5681. AAAI Press.

Zhen Zhang, Yuhua Zhao, Hang Gao, and Mengting Hu. 2024. Linkner: Linking local named entity recognition models to large language models using uncertainty. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4047–4058. ACM.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Appendix

A Datasets

Statistics	GMNER			FMNERG		
	Train	Valid	Test	Train	Valid	Test
Number	7000	1500	1500	7000	1500	1500
Entity	11782	2453	2543	11779	2450	2543
Groundable Entity	4694	986	1036	4733	991	1046
Box	5680	1166	1244	5723	1171	1254

Table 4: Data statistics across the GMNERG and FMNERG datasets.

We have compiled statistics for the GMNER and FMNERG datasets, including the total number of data entries, the number of entities, the number of entities with corresponding visual regions, and the number of visual entities, as detailed in Table 4. Specifically, the GMNER dataset contains four categories, while the FMNERG dataset includes eight coarse-grained categories and fifty-one fine-grained categories.

B Baselines

To evaluate the proposed framework, we adopt multiple frameworks and methods for comparison. Below are descriptions of these baseline approaches:

- Text-only. Extracting text entities without corresponding visual entities. HBiLSTM-CRF (Huang et al., 2015) uses an LSTM to encode the text sequence, followed by a CRF layer to classify the token categories. Bert and Bert-CRF (Devlin et al., 2019) replace the former backbone model with BERT. T5 and BART (Lewis et al., 2020; Raffel et al., 2020) treat entity recognition as a sequence generation task, using their generative capabilities to predict entities along with their categories.
- EVG-based. Firstly, text entities are extracted using various multimodal named entity recognition methods. Subsequently, corresponding visual entities that have been identified through object detection methods are selected. Two target detection models, RCNN and VinVL, (Zhang et al., 2021b; Girshick, 2015) are utilized to extract potential visual entities. GVATT (Lu et al., 2018) uses visual embeddings to initialize the hidden states of an LSTM, integrating visual context into the text processing sequence. UMT (Yu et al., 2020) employs a multimodal transformer to fuse image and text features, enhancing the interaction between modalities for improved recognition accuracy. UMGF (Zhang et al., 2021a) uses a graph-based approach to fuse multi-level modality features, providing a structured way to integrate diverse information sources. ITA (Wang et al., 2022b) supplements the model with sample knowledge for knowledge augmentation, aiming to enrich the contextual understanding of the entities. MMT5 and BART (Lewis et al., 2020; Raffel et al., 2020) treat entity recognition as a multimodal sequence generation task. Utilizing their generative capabilities, they predict entities along with their categories, effectively leveraging both text and image inputs.
- Unified-Generative. Simultaneously extracting text entities and selecting corresponding visual entities identified through object detection methods. Tiger and H-Index (Wang et al., 2023; Yu et al., 2023) use a multimodal sequence generation approach to simultaneously generate text entities and corresponding visual tokens, effectively integrating text and image data for enhanced entity recognition.

C Implementation Details

We conducted all experiments using a single NVIDIA RTX 4090 GPU and in the PyTorch framework. For optimization, we utilized the AdamW optimizer (Loshchilov and Hutter, 2019) to minimize the loss function. We set $\alpha = \beta = 0.1$ for textual entity recognition and $\gamma = 0.2$ for filtering samples across all datasets. The learning rate was set to $5e-6$, and a linear scheduler was employed to control it. The maximum sentence input length was capped at 256, and the mini-batch size was set to 4. The model underwent training for a total of 10 epochs. Additionally, We employed LoRA with the rank $= 64$ to instruction-tune LLaVA (Liu et al., 2023) for reranking within the $top5$ categories, with a learning rate of $5e-6$ over three epochs. We also fine-tuned BLIP (Li et al., 2022) with a learning rate of $5e-5$ for one epoch.

D Different LLMs for Span Prediction

We compared the effectiveness of knowledge augmentation in different LLMs in assisting with textual entity span prediction, as shown in Table 5. The performance of span prediction significantly improves with the assistance of any LLM, indicating that using LLMs as knowledge suppliers enables models to effectively capture phrases outside the vocabulary. Furthermore, the more common knowledge integrated into the LLM, the better its recognition performance.

Models	GMNER			FMNERG		
	Pre	Rec	F1	Pre	Rec	F1
-	87.01	87.43	87.22	87.24	87.58	87.41
LLaMA2-7B	87.62	88.03	87.82	87.58	87.99	87.78
LLaMA3-8B	87.91	88.25	88.08	87.11	89.03	88.06
ChatGPT	87.10	89.78	88.42	86.67	89.61	88.12

Table 5: Performance comparison across different LLMs on entity span prediction.

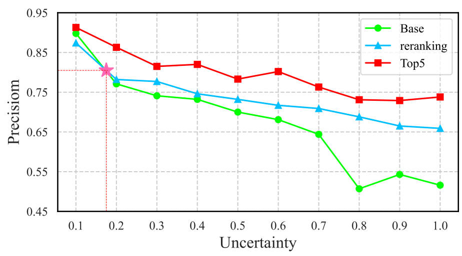

Figure 5: Prediction accuracy across varying levels of uncertainty in different settings.

E The Threshold for the Sample Filter

We explore the trend in which the precision of entity classification and the precision of the Top 5 categories vary with increasing uncertainty, and how the reranking model adjusts to identify the optimal threshold.

As shown in Figure 5, we observe a relatively clear trend: as the uncertainty of the predicted entity increases, the precision of entity classification decreases significantly. For the MLLM-based reranking model, this decline is more gradual, indicating that the MLLM performs better with difficult samples. We select the approximate value of γ where the precision levels of the reranking model and the base model converge as the threshold to filter samples.

F The number of candidates

We evaluate our model with different numbers of candidate categories, denoted as K. As shown in Figure 6, results across various models indicate that $K = 5$ yields the best performance. When K decreases, the probability of the ground truth being among the candidates diminishes. Conversely, as K increases, the inclusion of too many candidates makes it challenging for the model to select the most accurate prediction.

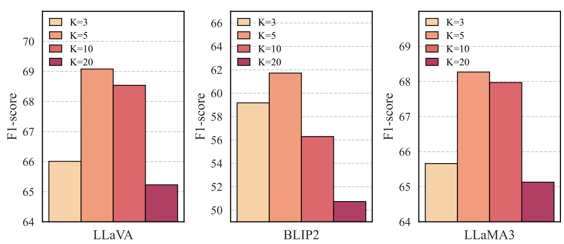

Figure 6: Performance comparison across various values of K for candidate categories.

G Prompt template

We present the template for various instructions used at different stages of our process. In Table 6, the GEM prompts the LLM by integrating text with the image caption to acquire the necessary knowledge. Notably, to acquire a stable format output, we opt for JSON as the output format. In Table 7, we pass the text and the image, label the entity with a special sign, and provide candidate categories to prompt our reranking model to select the best category. In Table 8, we simply concatenate the entity, given instruction, and category to feed into the grounding model, which then assesses the relevance between the text entity and the image.

Example 1
Query: Here's a tweet content: Details on our newly announced Federation Business Data Lake Solution :. Accordingly, it is corresponding caption: a screen shot of the federation business data lake 10. Please explain the tweet and especially some of the words in it according to your knowledge and answer in the format of json like: "explanation": " " **Answer:** The tweet is announcing a new solution called Federation Business Data Lake. The solution is likely designed to help businesses store and manage their data in a central repository, known as a data lake. A data lake is a large storage system that allows for the collection, storage, and analysis of large amounts of structured and unstructured data from various sources.
Example 2
Query: Here's a tweet content: i gotta feeling by the black eyed peas blocked you belong with me by taylor swift. Accordingly, it is corresponding caption: a screenshot of the youtube app showing the different videos. Please explain the tweet and especially some of the words in it according to your knowledge and answer in the format of json like: "explanation": " " **Answer:** The tweet is referencing the songs 'I Gotta Feeling' by The Black Eyed Peas and 'You Belong With Me' by Taylor Swift being blocked on YouTube. 'I Gotta Feeling' is a popular dance-pop song released in 2009, while 'You Belong With Me' is a country-pop song released in 2008. The tweet suggests that the user may have tried to access these songs on YouTube but found that they were blocked.
Example 3
Query: Here's a tweet content: Leonardo DiCaprio and Kate Moss on their way to an art show in New York City(1993) Accordingly, it is corresponding caption: a man and woman standing in the middle of a city street. Please explain the tweet and especially some of the words in it according to your knowledge and answer in the format of json like: "explanation": " " **Answer:** The tweet is referencing a photograph of Leonardo DiCaprio and Kate Moss attending an art show in New York City in 1993. Leonardo DiCaprio and Kate Moss are both well-known actors and models. In the image, a man and woman (presumably Leonardo DiCaprio and Kate Moss) are seen standing in the middle of a city street, likely on their way to the art show. The tweet is highlighting a moment from their past where they attended a cultural event together.

Table 6: Templates for knowledge augmentation module.

Example 1
Query:
Here's a tweet along with its image, where the entities of the text have been labeled using &&, and you need to determine his category from the candidate set.
Text: The geometry of plants . Garfield(&&) Park(&&) Conservatory.
Candidate: ["park", "continent", "city", "country", "software"].
Answer:
park
Example 2
Query:
Here's a tweet along with its image, where the entities of the text have been labeled using &&, and you need to determine his category from the candidate set.
Text: Golden(&&) State(&&) Warriors(&&) win NBA championship against Cleveland Cavaliers.
Candidate: ["company", "sports_team", "sports_league", "magazine", "social_organization"].
Answer:
sports_team
Example 3
Query:
Here's a tweet along with its image, where the entities of the text have been labeled using &&, and you need to determine his category from the candidate set.
Text: RT @ AwkwardGoogle : Harry(&&) Potter(&&).
Candidate: ["author", "character", "coach", "event_other", "actor"].
Answer:
character

Table 7: Templates for reranking module.

Example 1: Cleveland is belong sports_team.
Example 2: taylor swift is belong musician.
Example 3: The Edge of the Sea is belong written_work.

Table 8: Templates for grounding module.