UnCo: Uncertainty-Driven Collaborative Framework of Large and Small Models for Grounded Multimodal NER

Jielong Tang[1,3], **Yang Yang**[2], **Jianxing Yu**[1,3], **Zhenxing Wang**[4],
Haoyuan Liang[1], **Liang Yao**[2], **Jian Yin**[1,3,*]

[1]School of Artificial Intelligence, Sun Yat-sen University
[2]School of Cyber Science and Technology, Sun Yat-sen University
[3]Key Laboratory of Sustainable Tourism Smart Assessment Technology, Ministry of Culture and Tourism
[4]Institute of Software, Chinese Academy of Sciences

{tangjlong3,yangy2233,lianghy68}@mail2.sysu.edu.cn, {yujx26,yaoliang3,issjyin}@mail.sysu.edu.cn,
wangzhenxing@iscas.ac.cn

Abstract

Grounded Multimodal Named Entity Recognition (GMNER) is a new information extraction task. It requires models to extract named entities and ground them to real-world visual objects. Previous methods, relying on domain-specific fine-tuning, struggle with unseen multimodal entities due to limited knowledge and generalization. Recently, multimodal large language models (MLLMs) have demonstrated strong open-set abilities. However, their performance is hindered by the lack of in-domain knowledge due to costly training for GMNER datasets. To address these limitations, we propose **UnCo**, a two-stage **Un**certainty-driven **Co**llaborative framework that leverages the complementary strengths of small fine-tuned models and MLLMs. Specifically, **in stage one**, we equip the small model with a unified uncertainty estimation (UE) for multimodal entities. This enables the small model to express *"I do not know"* when recognizing unseen entities beyond its capabilities. Predictions with high uncertainty are then filtered and delegated to the MLLM. **In stage two**, an Uncertainty-aware Hierarchical Correction mechanism guides the MLLM to refine uncertain predictions using its open-domain knowledge. Ultimately, UnCo effectively retains the in-domain knowledge of small models while utilizing the capabilities of MLLMs to handle unseen samples. Extensive experiments demonstrate UnCo's effectiveness on two GMNER benchmarks.

1 Introduction

Grounded Multimodal Named Entity Recognition (GMNER) is a pivotal task in multimodal information extraction, aiming to identify textual named entities and their corresponding visual regions within image-text data. This task holds significant promise for various downstream applications such as multimodal knowledge graph construction (Liu et al.,
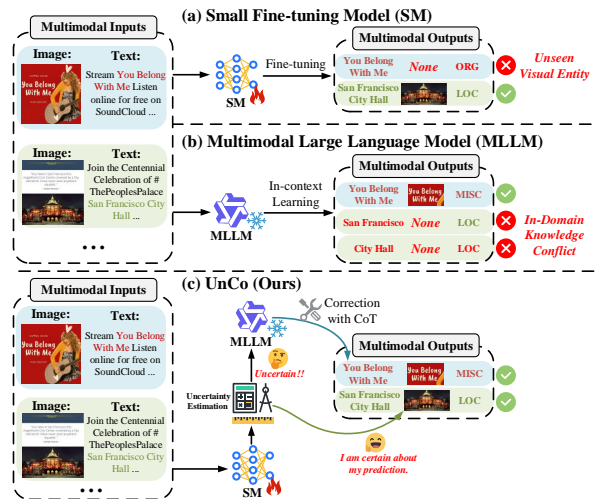

Figure 1: The comparison of existing approaches and our **UnCo** framework: (a) The small model can learns in-domain knowledge but miss unseen visual entities. (b) The MLLM can detect unseen samples but lack of in-domain knowledge. (c) Our **UnCo** combines the complementary strengths of both models via uncertainty estimation (UE).

2019), VQA (Li et al., 2025), etc. Previous methods (Li et al., 2024a) typically train a sequence labeling model to extract textual entities, followed by a visual grounding model to identify related visual regions. To address error propagation, some studies (Yu et al., 2023; Wang et al., 2023a; Tang et al., 2025) focus on generating span-type-region triplets by fine-tuning end-to-end transformer models on annotated GMNER datasets.

However, grounding named entities to real-world visual objects inherently poses an *open-world* challenge (Ren et al., 2024). These fine-tuned models struggle to recognize previously *unseen multimodal entities* due to their limited knowledge and lack of generalizable visual capabilities, such as fine-grained object detection (Wang et al., 2024b), optical character recognition (OCR) (Wang et al., 2022b), and scene graph understanding (Zhang

---

*Corresponding Authors.

et al., 2021a). As illustrated in Figure 1 (a), small fine-tuned models[1] fail to identify both the visual region and entity type for *You Belong With Me*. This failure stems from two primary limitations: (1) the model's inability to recognize the image as an album cover poster, which requires background knowledge, and (2) its lack of OCR capability to detect the text left of the central image as salient entity regions. Consequently, only fine-tuning a small model is insufficient for the GMNER task.

Recent advances in multimodal large language models (MLLMs) have demonstrated impressive cross-modal capabilities in CV and NLP domains. These models possess extensive multimodal knowledge and effectively leverage it for open-world visual perception, making them a promising solution for the GMNER task. However, applying MLLMs to extract multimodal entities still presents challenges. First, training task-specific MLLMs for GMNER is computationally intensive and may lead to catastrophic forgetting of pre-trained knowledge (Luo et al., 2023). In addition, in context-learning paradigms (Min et al., 2022) often yield suboptimal performance due to knowledge discrepancies between MLLMs and domain-specific dataset, which we term as *In-Domain Knowledge Conflict*. As shown in Figure 1 (b), *San Francisco City Hall* is labeled as a *LOC (Location)* entity in the social media dataset, while the MLLM splits it into two different entities: *San Francisco* and *City Hall*. In contrast, small models can effectively capture this in-domain knowledge through fine-tuning.

Given these observations, we maintain that neither small models nor MLLMs alone can effectively address GMNER challenges. Motivated by this, we propose **UnCo**, a two-stage **Un**certainty-driven **Co**llaborative framework that integrates the complementary strengths of small models and MLLMs, as shown in Figure 1 (c). Concretely, in **Stage One**, small models initially predict multimodal entity triplets based on their in-domain knowledge and identify those hard samples beyond their capacity. To achieve this, we first fine-tune a small autoregressive model (Lewis et al., 2019) to generate structured multimodal triplets. Inspired by previous studies (Kendall and Gal, 2017) that models often produce unstable outputs when predictions are incorrect, we introduce a unified Uncertainty Estimation (UE) module using multiple

Monte Carlo dropout sampling (Gal and Ghahramani, 2016). This UE module assigns higher uncertainty to difficult or unseen entities. Additionally, since textual entities are not always present in images, this inter-modality inconsistency will impact UE performance (Jung et al., 2023). To address this, we propose a modality representation debiasing module to enhance UE's robustness. In **Stage Two**, the MLLM is activated when uncertainties exceed a predefined threshold. However, due to the inherent hallucination issues in MLLMs (Bai et al., 2024), directly generating multimodal predictions via MLLM does not achieve optimal performance (see detailed analysis in Section 2.3). Moreover, attributes of entities predicted by the small model can provide auxiliary information for MLLM predictions. To this end, we propose the Uncertainty-aware Hierarchical Correction mechanism, which uses pre-detected entities and their associated uncertainty scores as key indicators to guide MLLM in iteratively refining entity attributes. Ultimately, results from both the small and MLLM are integrated for the final prediction. In this way, **UnCo** preserves the in-domain knowledge of the small model while using MLLMs to generalize to unseen multimodal entities. Our contributions are summarized as follows:

- We propose a novel collaborative framework of large and small models, named UnCo, for the GMNER task, which leverages the powerful general capabilities of MLLMs to handle unseen multimodal entities while retaining the domain-specific knowledge learned by small fine-tuned models.

- We introduce a unified uncertainty estimation based on Monte Carlo Dropout for multimodal entities, along with a debiasing module to reduce modality representation inconsistency, achieving more robust uncertainty estimation performance.

- We validate several mainstream MLLMs in UnCo and conduct extensive experiments on two benchmarks, demonstrating that our method outperforms existing state-of-the-art (SOTA) models.

2 Our Method

2.1 Overview

Task Formulation. Given a sentence X and its corresponding image I, the Grounded Mul-

---

[1]In this paper, small fine-tuned models refers to models specifically adapted to a narrow dataset, focusing on particular tasks or domains.
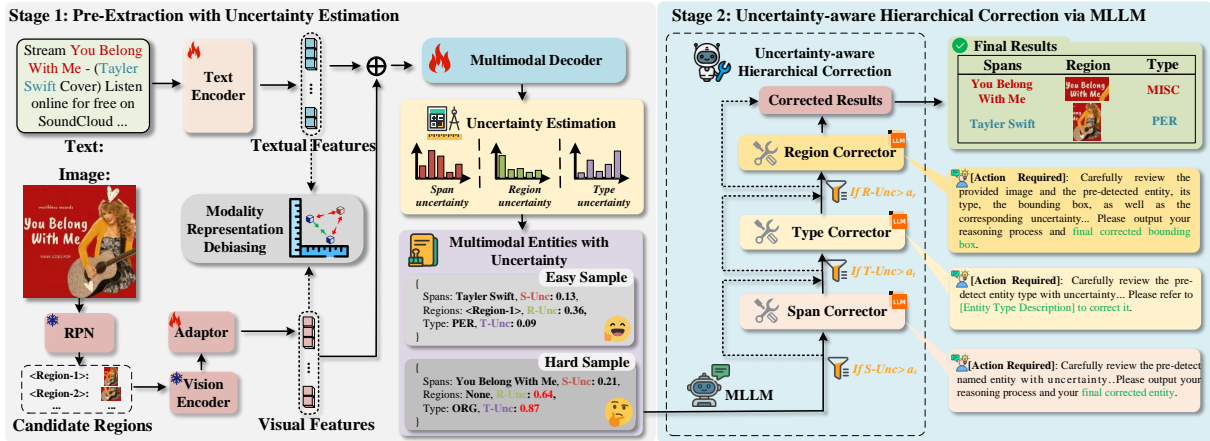
Figure 2: Overview of our UnCo. **In stage 1**, we first obtain pre-detected multimodal entities using a small fine-tuned model equipped with uncertainty estimation. **In stage 2**, these initial predictions are then refined through an uncertainty-aware hierarchical correction mechanism leveraging a multimodal large language model.

timodal Named Entity Recognition (GMNER) task aims to predict a set of multimodal triplets $\{(e_i^s, e_i^t, e_i^r)\}_{i=1}^m$, where e_i^s is the i-th entity span, e_i^t denotes its type, and e_i^r represents the bounding box coordinates of the entity in image. To allow the model to output uncertainties for predictions, we introduce parallel triplets $\{(u_i^s, u_i^t, u_i^r)\}_{i=1}^m$, where u_i^s, u_i^t, u_i^r denote entity span, type, and region uncertainties respectively.

Overall Workflow. As shown in Figure 2, our **UnCo** contains two stages: **In Stage 1**, a small fine-tuned GMNER model pre-extracts multimodal entity triplets, while an dropout layer is activated to quantify prediction uncertainties during inference period. **In Stage 2**, the multimodal large language model (MLLM) performs progressive correction of predicted triplets guided by uncertainty signals. The final prediction is obtained by integrating the outputs from both the local model and the MLLM.

2.2 Stage 1: Pre-Extraction with Uncertainty Estimation

The inherent diversity of visual entities makes it challenging for previous models to reliably detect unseen visual entities during inference. To address this, we first introduce a sequence-to-sequence model to simultaneously generate multiple multimodal entity triplets. After that, a unified uncertainty estimation (UE) is proposed for these triplets, enabling the model to express *"I do not know"* when facing uncertain predictions, and then delegate them to MLLMs for further refinement. The preliminary knowledge for uncertainty estimation is provided in Appendix A.

GMNER Modeling. Prior works (Li et al., 2024a) decompose GMNER into MNER and entity grounding with isolated models, but this paradigm struggles to achieve consistent uncertainty estimation. To address this, we formulate GMNER as a sequence generation task, where entity span, type, and region uncertainties are inherently transformed to unified token-level uncertainties in the target sequence. Following (Yu et al., 2023), we employ a sequence-to-sequence model BART (Lewis et al., 2019) as our backbone. The original text is embedded as $H_T = (h_t^1, ..., h_t^S)$. For visual representation, we first utilize freezing VinVL (Zhang et al., 2021b) as region proposal network (RPN) to generate K candidate regions. These candidate regions, along with the entire image, are fed into the pre-trained vision transformer (Dosovitskiy et al., 2020) to obtain regional visual tokens. Besides, a visual adaptor (MLP) is trained to map visual features into the same dimension and semantic space as textual features. The processed visual representation is denoted as $H_V = (h_r^1, ..., h_r^{K+1})$, where h_r^{K+1} is the whole image representation. Finally, the concatenation of visual and textual representation is fed to the decoder.

In the decoding phase, the model autoregressively generates multimodal triplets in template $(</s>, e_i^s, </r>, e_i^t)$ one-by-one, where $</s>$ and $</r>$ are special tokens for *start of entity* and *region indicator*[2], respectively. The objective of genera-

---

[2] Different entity types, $(</s>)$, and $(</r>)$ are embedded as special tokens in the vocabulary.

tive model can be formalized as:

$$H_t = \text{Decoder}(y_{<t}, [H_T; H_V]),$$

$$p(\boldsymbol{y}_t) = \text{Softmax}(W_1^T H_t),$$

$$\mathcal{L}_{token} = -\frac{1}{NL} \sum_{j=1}^{N} \sum_{t=1}^{L} \log p(\boldsymbol{y}_t^j) \qquad (1)$$

where L is the length of the output sequence, W_1 is the linear transformation matrix. For supervised entity grounding, we first obtain the hidden states of `</r>`, denoted as H_k, and then calculate the probability distribution over all the candidate regions:

$$p(\boldsymbol{g}_k) = \text{Softmax}(W_2^T H_k) \qquad (2)$$

where $p(\boldsymbol{g}_k) \in \mathbb{R}^{1 \times (K+1)}$ is the probability of regions matching, the $(K+1)^{th}$ region is used to matching those ungroundable entities[3]. Following (Yu et al., 2023; Wang et al., 2023a), we compute the Intersection over Union (IoU) scores between candidate regions and ground truth regions. IoU of the $(K+1)^{th}$ region will be set to 1 when the entity is ungroundable; otherwise 0. The Kullback-Leibler Divergence (KLD) loss will be used to optimize model's parameters:

$$\mathcal{L}_{region} = \frac{1}{NM} \sum_{j=1}^{N} \sum_{k=1}^{M} IoU_k^j \log \frac{IoU_k^j}{p(\boldsymbol{g}_k^j)}, \qquad (3)$$

where IoU is the normalized IoU score distribution of supervised region grounding, M is the number of entities.

Unified Uncertainty Estimation for Multimodal Entity. In this paper, we adopt sampling-based Monte Carlo Dropout (MCD) (Gal and Ghahramani, 2016) to quantify uncertainty. As a training-free approach, MCD offers seamless integration with existing GMNER models. Given the trained sequence-to-sequence model f_θ, MCD approximates Bayesian inference by performing T stochastic forward passes with different dropout masks. To reduce the computational cost, we only sample a random dropout mask matrix $\mathbf{m}^{(i)}$ in the decoder. At each timestep t, token-level logits sampling can be represented as:

$$H_t^{(i)} = f_\theta(y_{<t}, [H_T; H_V], \mathbf{m}^{(i)}),$$

$$\mathbf{m}^{(i)} \sim \text{Bernoulli}(p), \quad i = 1, \dots, T \qquad (4)$$

where $p \in [0, 1]$ is the dropout rate, and T is the number of MC forward. The token-level uncertainty u_t^{token} can be calculated as the entropy of MCD integration[4]:

$$p_c^{(i)} = \text{Softmax}(W^T H_t^{(i)}) \qquad (5)$$

$$u_t^{token} = -\sum_{c \in \mathcal{V}} \left(\frac{1}{T} \sum_{i=1}^{T} p_c^{(i)} \right) \log \left(\frac{1}{T} \sum_{i=1}^{T} p_c^{(i)} \right) \qquad (6)$$

where $\mathcal{V}$ is the label set. Similarly, the entity grounding uncertainty can be calculated using Eq.(5) and Eq.(6), denoted as u_t^{region}. Finally, the i-th multimodal entity uncertainty unc_i in span, type, and region are formalized as follow:

$$unc_i : \begin{cases} u_i^s = \dfrac{\sum_t u_t^{token}}{|e_i^s|}, & t \text{ in } (\text{</s>} ... \text{</r>}), \\ u_i^t = u_t^{token}, & t \text{ after } \text{</r>}, \\ u_i^r = u_t^{region}, & t \text{ in } \text{</r>} \end{cases} \qquad (7)$$

where $|e_i^s|$ denotes the length of the predicted entity span.

Modality Representation Debiasing. Since most visual candidate regions cannot be matched with textual entities, direct fusion of visual and textual features introduces inter-modality inconsistency that makes model's uncertainty estimation difficult. To mitigate this, we propose a dynamic debiasing module for cross-modal representations. Specifically, we obtain the textual entity representation h_{ent}^i via average pooling of its token features. h_{ent}^i acts as the anchor to select positive or negative regions pairs: positive if IoU > 0.5, otherwise negative. Considering that some negative samples still contain partial visual entity information due to overlap with the ground truth, we use IoU scores as a dynamic margin in the triplet ranking loss (Schroff et al., 2015) to serve as a debiasing objective:

$$\mathcal{L}_{debiasing} = \sum_i \sum_{j \neq i^+} \max \bigg(0, \, D(h_{\text{ent}}^i, h_{\text{r}}^{i+})$$
$$- D(h_{\text{ent}}^i, h_{\text{r}}^j) + \lambda(1 - \text{IoU}_{ij}) \bigg) \qquad (8)$$

where h_{r}^{i+} refers to positive samples $D(a, b) = 1 - \frac{a \cdot b}{\|a\| \|b\|}$ is a distance function, λ is a hyperparameter. This formulation ensures that entity-region

---

[3]Ungroundable means entities do not exist in the image or their regions are not included in candidate regions.

[4]The largest entropy value occurs when all token labels have the same probability. It reflects model is not sure about its predictions (Zhang et al., 2024).

Methods	GMNER	MNER	EEG
Baseline (small)	56.88	79.27	61.75
Qwen2.5VL-72B (Bai et al., 2025)			
Generation-based	55.91	78.52	60.95
Correction-based (UnCo)	62.14	81.33	66.72

Table 1: Empirical study of prompting MLLM on Twitter-GMNER dataset (Yu et al., 2023). Generation-based indicates directly generating span-type-region triplets for filtered samples. Correction-based represents correcting pre-extracted results of small model.

pairs with lower IoU receive proportionally larger distance penalties. Finally, the overall training objective is:

$$\mathcal{L}_{overall} = \mathcal{L}_{token} + \mathcal{L}_{region} + \gamma\mathcal{L}_{debiasing} \quad (9)$$

where γ is a hyperparameter.

2.3 Stage 2: Uncertainty-aware Hierarchical Correction via MLLM

When we have obtained the entities predicted by the small model and their corresponding types, regions, and uncertainties, we start the refinement process. First, we filter out uncertain samples using predefined thresholds. An intuitive approach to refine these samples is to directly generate their span-type-region triplets by prompting MLLM. However, as shown in Table 1, the empirical results reveal that generation-based method does not improve GMNER performance. The reason is that due to the inherent hallucination of MLLM, more noisy entity attributes are mistakenly generated. In contrast, the correction-based method significantly outperforms the generation-based approach. This is due to pre-extracted entity attributes and uncertainty can provide auxiliary information to the large model, alleviating its hallucination issues. For example, the pre-detected 4D bounding boxes of entities serve as visual prompts for the MLLM, while uncertainty indicates the intensity of error. Correcting an existing entity region is much easier than generating a new one. To this end, we propose the Uncertainty-aware Hierarchical Correction mechanism to refine the pre-detected entities.

Specifically, we define three different correctors, namely the Span Corrector, Type Corrector, and Region Corrector. The prompts for these three correctors are constructed according to pre-designed template (Please refer to Appendix D for details). We define the prompt template as four parts, namely

---

Algorithm 1: Uncertainty-aware Hierarchical Correction

Input: Entity set $E = \{(e_i^s, e_i^t, e_i^r)\}_{i=1}^m$, uncertainties set $\{(u_i^s, u_i^t, u_i^r)\}_{i=1}^m$, and uncertainty thresholds $\theta_s, \theta_t, \theta_r$

Output: Corrected entity set $E = \{(\hat{e}_i^s, \hat{e}_i^t, \hat{e}_i^r)\}_{i=1}^m$

1 **for** *entity i in E* **do**
2   **if** $u_i^s > \theta_s$ **then**
3   Call SPANCORRECTOR to refine e_i^s based on u_i^s;
4   **end**
5   Update $\hat{e}_i^s$ in E
6   **if** $u_i^t > \theta_t$ **then**
7   Call TYPECORRECTOR to refine e_i^t, based on u_i^t and $\hat{e}_i^s$;
8   **end**
9   Update $\hat{e}_i^t$ in E
10   **if** $u_i^r > \theta_r$ **then**
11   Call REGIONCORRECTOR to refine e_i^r, based on u_i^r, $\hat{e}_i^s$, and $\hat{e}_i^s$;
12   **end**
13   Update $\hat{e}_i^r$ in E
14 **end**
15 **return** E

---

role definition, entity type definition (if needed), output format definition, and action definition. In addition, we introduced prompt-based CoT (Wei et al., 2022) in the action definition to guide MLLM to think during the correction process, thus yielding more accurate results. Under the guidance of uncertainty, we perform corrections step by step. The detail is shown in Algorithm 1.

3 Experiments

3.1 Experiment Setting

Datasets. We conduct extensive experiment on two benchmarks: Twitter-GMNER (Yu et al., 2023) and Twitter-FMNERG (Wang et al., 2023a). Details of two datasets are in Appendix B.3.

Evaluation. Based on (Yu et al., 2023), we evaluate GMNER and its two subtasks: Multimodal Named Entity Recognition (MNER) and Entity Extraction & Grounding (EEG). MNER focuses on identifying entity spans and types, while EEG targets extracting entity spans and regions. We use F1 Score as metrics for GMNER and its subtasks. To assess uncertainty estimation, we employ the Area Under the Receiver-Operator Curve (AUROC) (Hu

Methods	Twitter-GMNER			Twitter-FMNERG		
	GMNER	MNER	EEG	GMNER	MNER	EEG
GPT4o (Hurst et al., 2024)	41.29	65.07	44.95	32.37	52.26	41.60
GVATT-OD-EVG (Lu et al., 2018a)	48.57	76.26	53.32	40.32	60.35	54.35
UMT-OD-EVG (Yu et al., 2020)	50.29	78.58	54.78	41.32	61.63	54.43
UMGF-OD-EVG (Zhang et al., 2021a)	51.67	78.83	55.74	41.92	61.79	54.75
ITA-OD-EVG (Wang et al., 2022b)	51.56	79.37	55.69	42.78	63.21	57.26
MMT5 / BARTMNER-OD-EVG (Yu et al., 2023)	52.45	80.39	55.66	45.21	66.61	58.18
H-Index (Yu et al., 2023)	56.41	79.73	61.18	46.55	64.84	60.46
TIGER (Wang et al., 2023a)	57.48	-	-	47.20	64.91	61.96
GMDA† (Li et al., 2024b)	58.61	-	-	47.37	-	-
MQSPN (Tang et al., 2025)	58.76	80.43	62.40	47.86	66.83	61.95
GEM† (Wang et al., 2024b)	59.83	**83.15**	63.19	50.54	**68.09**	63.59
RiVEG† (Li et al., 2024a)	<u>63.80</u>	<u>82.89</u>	66.92	-	-	-
Our Baseline (Small Model)	56.88	79.27	61.75	46.79	64.78	61.44
InternVL3-9B (Zhu et al., 2025)	34.10	60.46	36.26	24.13	41.54	34.37
-w / Direct Correction	46.83	73.52	53.37	28.58	47.39	41.94
-w / UnCo	58.79	80.09	62.95	47.92	65.33	62.30
Qwen2.5VL-7B (Bai et al., 2025)	35.29	57.51	39.93	23.56	39.69	36.83
-w / Direct Correction	45.20	72.06	53.92	27.51	45.35	42.22
-w / UnCo	58.83	79.55	63.49	48.17	65.06	62.73
Qwen2.5VL-72B (Bai et al., 2025)	40.74	62.18	46.01	31.25	48.62	42.29
-w / Direct Correction	50.69	75.26	57.53	40.82	58.56	56.37
-w / UnCo	62.14	81.33	<u>66.72</u>	<u>52.44</u>	67.21	<u>65.68</u>
Gemini-2.5 Pro (Comanici et al., 2025)	43.31	64.57	47.63	34.02	51.14	45.89
-w / Direct Correction	53.54	75.75	60.09	42.49	61.62	58.33
-w / UnCo	**64.58**	81.71	**69.62**	**53.56**	<u>67.70</u>	**68.25**

Table 2: Comparisons of various competitive approaches on two GMNER datasets are presented. Bold text indicates the best result, while underlined text denotes the second-best. The results of MLLMs baselines are based on a 3-shot In-context Learning. † indicates the methods using additional data or knowledge augmentation.

Components			Twitter-GMNER			Twitter-FMNERG	
(a) UE	(b) **MRD**	(c) UHC	GMNER	MNER	EEG	MNER	EEG
(I) Baseline			56.88	79.27	61.75	64.78	61.44
(II) Qwen2.5VL-72B			40.74	62.18	46.01	48.62	42.29
Ablation for only small fine-tunded model of UnCo							
✓	✗	-	56.94	79.45	61.67	64.81	61.40
✗	✓	-	57.33	79.60	61.89	64.93	61.68
✓	✓	-	57.54	79.81	61.91	65.02	61.87
Ablation for full UnCo							
✓	✓	✗	58.91	79.52	62.95	65.31	62.56
✓	✗	✓	60.25	80.76	65.03	66.49	64.20
✗	✓	✓	53.08	75.92	59.17	57.68	56.52
✓	✓	✓	**62.14**	**81.33**	**66.72**	**67.21**	**65.68**

Table 3: Ablation study results. **(a)** Uncertainty Estimation (UE). **(b)** Modality Representation Debiasing (MRD). **(c)** Uncertainty-aware Hierarchical Correction (UHC).

et al., 2023) for distinguishing incorrect from correct predictions. Descriptions of the evaluation metrics, baseline methods and implementation details can be found in Appendix B.4, B.2, and B.1, respectively.

3.2 Performance Comparison

Comparison with State-of-the-arts. Rows 2–13 in Table 2 compare UnCo with a range of competitive baselines. All variants of UnCo consistently outperform both our small-model baseline and all fine-tuned models across the two datasets. Unlike approaches that rely on extensive augmentation and additional supervision, UnCo achieves substantial gains without extra training. Moreover, UnCo remains competitive with state-of-the-art systems such as RiVEG and GEM, where RiVEG leverages a stronger backbone (OFA-large) and LLM-based knowledge augmentation, and GEM fine-tunes two MLLMs (LLaVA and BLIP2). In contrast, UnCo (Gemini-2.5 Pro) surpasses them, delivering improvements of 0.78% on Twitter-GMNER and 3.02% on Twitter-FMNERG. Most of these gains stem from more accurate extraction of visual entities, yielding 2.70% and 4.66% improvements on the EEG task. This reflects the higher variability of visual entities in Twitter data, which presents an open-world challenge. By combining the open-set generalization and knowledge capacity of MLLMs with the domain-specific knowledge of supervised models, UnCo effectively bridges this gap.

Comparison with MLLMs. As shown in Table 2 rows 14 to 31, we conducted experiments on UnCo with four different MLLMs: Qwen2.5VL, internVL3, GPT4o, and Gemini 2.5 Pro to verify its scalability. Furthermore, a pipeline named *Direct Correction* is designed as comparative methods, where MLLMs directly correct all pre-detected entities. The results demonstrate that UnCo signif-
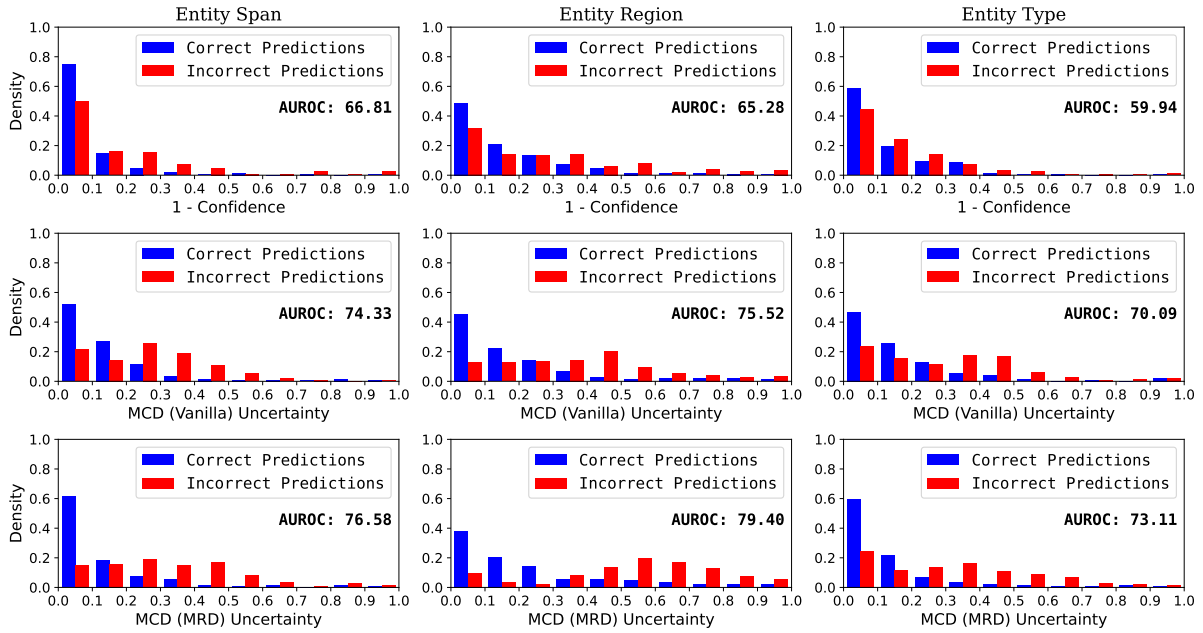
Figure 3: The density distribution of correct and incorrect predictions under different uncertainty intervals (normalized to [0,1]). Higher AUROC values indicate better filtering of incorrect samples. MCD (Vanilla) uses only MC Dropout in UnCo, while MCD (MRD) includes additional training for modality representation debiasing.
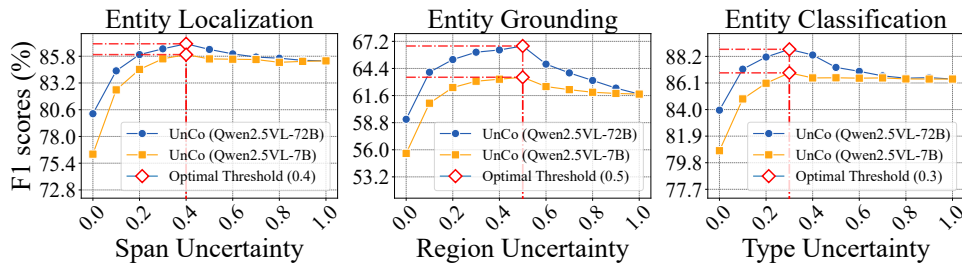

Figure 4: Performance under different uncertainty for entity localization, grounding and classification tasks. Thresholds are selected in validation set.

icantly outperforms both the in-context learning of MLLMs and the *Direct Correction* pipeline because: (1) Few-shot in-context learning in MLLMs struggles to acquire sufficient domain-specific knowledge, while UnCo can solve it with small model. (2) Compared to the *Direct Correction* pipeline, UnCo's uncertainty effectively filters out-of-domain entities, enabling MLLMs to focus on correcting uncertain entities. In contrast, directly correcting all entities will lead to erroneous modifications in originally accurate predictions, primarily due to knowledge discrepancy between MLLMs and GMNER datasets.

3.3 Ablation Studies

In this section, we conduct a series of ablation studies to validate the contributions of different components on the GMNER task and its subtasks.

The experimental results are presented in Table 3. **(1) Effectiveness of Uncertainty Estimation (UE).** Removing the UE module leads to the a large performance declines in UnCo across all tasks, by 9.06% in the Twitter GMNER main task and 9.53% in the Twitter-FMNERG's MNER subtask. This demonstrates that the UE module can effectively recognize incorrectly predicted entities, enhancing model's performance. **(2) Effectiveness of Modality Representation Debiasing (MRD).** Excluding the MRD results in performance reductions of 1.89%, 0.57%, and 1.69% in the Twitter GMNER and its MNER and EEG subtasks, respectively. Coupled with results from Figure 3, it reveals that improved modality representation enhances the robustness of uncertainty estimation in multimodal tasks. **(3) Effectiveness of Uncertainty-aware Hierarchical Correction (UHC).** In row 10 of Table

7651

3, we replace the UHC module by MLLM one-time correction. This results in a performance drop of 3.23% in GMNER. This is because the GMNER task involves three subtasks: entity localization, classification, and grounding. Some decisions on entity attributes require iterative reasoning that integrates multiple sources of knowledge and visual cues. Directly outputting correction results can easily lead to hallucinations in MLLMs, thereby reducing overall performance. More ablation analysis on UHC are illustrated in Appendix C.1.

3.4 Analysis and Discussion

Can UnCo filter out incorrect entities? Figure 3 shows density distributions and AUROC values for correct and incorrect entity elements across methods. Initially, raw confidence struggles to distinguish incorrect entities because softmax probabilities tend to make all predictions overconfident. Adding Monte Carlo Dropout smooths the incorrect sample distribution and significantly boosts AUROC. Further inclusion of the MRD module enhances AUROC by 2.25%, 3.88%, and 3.02% for entity span, region, and type. These results demonstrate the effectiveness of MCD and MRD on filtering incorrect predictions.

Analysis of Uncertainty Thresholds. From previous results and analysis, we observe that uncertainty plays a crucial role in UnCo. To determine the optimal thresholds, we conduct tuning experiments on three uncertainties (span, region, type) using the Twitter GMNER validation set, as shown in Figure 4. Firstly, setting the uncertainty threshold too low (i.e., threshold < 0.1) allows many correct entities to be processed by MLLMs, leading to performance degradation. This is primarily due to MLLMs' in-domain knowledge conflict and hallucination issues, which cause originally correct entities to be mismodified. Conversely, a high uncertainty threshold (i.e., threshold > 0.8) filters most entities, resulting in minimal performance gains. For Twitter GMNER, the optimal thresholds for span, region, and type are 0.4, 0.5, and 0.3, respectively.

Case Study. We conduct a comprehensive case study on the Twitter-GMNER test set, as shown in Figure 5. In the first stage, UnCo (Small) generates pre-detected results with uncertainty scores. We observe that the visual regions corresponding to the entities *Antoine Vermette* and *Ben Bishop* are incorrectly identified. Due to the limited fine-grained visual understanding capabilities of the small model,
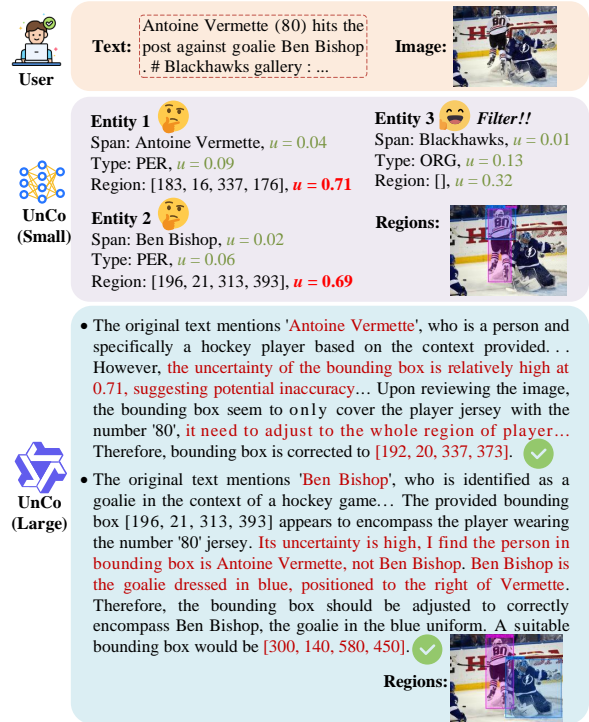

Figure 5: Predictions and Correction Process of UnCo.

it struggles to distinguish these PERSON entities in the image, resulting in high uncertainties. In the second stage, these uncertain results are refined by UnCo (Large). Under the guidance of uncertainty scores, the large model successfully capture visual cues like the *jersey number 80*, enabling it to differentiate the entities. For *Ben Bishop*, it leverages contextual understanding of *goalie* to produce the correct bounding box. This demonstrates the large model's superior reasoning and visual comprehension in resolving small model's weakness.

4 Related Work

Grounded Multimodal Named Entity Recognition (GMNER). Unlike traditional multimodal named entity recognition (Lu et al., 2018b; Zhang et al., 2018; Yu et al., 2020) solely detecting textual named entities, GMNER aims to extract multimodal entity information, including entity span, type, and corresponding visual regions from image-text pairs. It serves a wide range of downstream tasks, such as question answering systems (Yu et al., 2021, 2025) and knowledge bases (Wang et al., 2023b). Existing methods (Yu et al., 2023; Wang et al., 2023a; Li et al., 2024b; Tang et al., 2025) focus on detecting span-type-region triplets by fine-tuning transformer-based models (Lewis et al., 2019; Devlin et al., 2019) on GMNER

datasets. However, limited background knowledge and weak generalization capabilities hinder their performance, especially in visual entity grounding. Some studies attempt to enhance these models with external knowledge from search engines (Wang et al., 2022a; Ok et al., 2024) or large language models (Li et al., 2024a; Wang et al., 2024b; Liu et al., 2024). However, effectively leveraging this knowledge for entity grounding remains challenging due to restricted visual generalization. Unlike these approaches, our UnCo utilizes the knowledge and generalization capabilities of multimodal large language models to assist small fine-tuned models in handling unseen samples.

Interaction of Small and Large Models. Large language models (LLMs) exhibit strong capabilities but face challenges like hallucinations, intensive fine-tuning, and limited interpretability (Wang et al., 2024a). Using small language models (SLMs) to enhance LLMs is emerging as a new paradigm. (Azaria and Mitchell, 2023) employ a BERT classifier to evaluate the truthfulness of LLMs internal states, reducing hallucinations. SuperICL (Xu et al., 2024) and SuperContext (Yang et al., 2024) integrate predicted labels and confidence from SLMs to improve LLM performance and knowledge transfer. (Zhang et al., 2024) are the first to filter out-of-domain entities by small model uncertainty and then apply LLM for classification. Different from approaches that solely focus on unimodal data, our UnCo method enhances the extraction of multimodal entity information, including entity span, type, and region, effectively addressing more challenging tasks.

Conclusion

We present **UnCo**, a novel collaborative framework that synergizes the strengths of small fine-tuned models and multimodal large language models (MLLMs) to address the open-world challenges in GMNER. UnCo introduces a two-stage pipeline: (1) a small model generates entity triplets with unified uncertainty estimation (2) an uncertainty-aware hierarchical correction mechanism guides MLLMs to refine predictions progressively. Extensive experiments with diverse MLLMs demonstrate UnCo's effectiveness across benchmarks.

Limitations

GMNER remains a challenging task, particularly in real-world scenarios where it is necessary to ex-

tract open-world entities and localize them within specific visual regions. UnCo introduces a novel approach that leverages both MLLMs and small fine-tuned models to enhance performance for unseen multimodal entities. However, the selection of effective uncertainty thresholds requires hyperparameter tuning on a validation subset of the domain dataset, and such thresholds are typically domain-specific. Additionally, the inherent hallucinations and knowledge limitations of MLLMs pose challenges for GMNER applications in more specialized domains. In the future, exploring adaptive threshold selection algorithms and integrating external knowledge sources, such as domain-specific knowledge graphs or multimodal Retrieval-Augmented Generation (mRAG) techniques, could offer promising solutions to address these challenges.

Ethics Statement

The technology proposed in this paper enables the extraction of multimodal entity information. Our framework is built on a combination of a small fine-tuned model and a multimodal large language model (MLLM) for refining pre-detected entity predictions. However, since some closed-source MLLMs rely on online API calls, there is a risk of exposing private domain data to external servers, potentially leading to privacy concerns. To address this issue, we recommend using locally deployed MLLMs for processing private domain data, while reserving API-based methods for extracting information from public domain data.

Acknowledgements

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie

Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. 2023. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. *Advances in Neural Information Processing Systems*, 36:42191–42216.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Fan Li, Jianxing Yu, Jielong Tang, Wenqing Chen, Hanjiang Lai, Yanghui Rao, and Jian Yin. 2025. Answering complex geographic questions by adaptive reasoning with visual context and external commonsense knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25498–25514.

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

Ziyan Li, Jianfei Yu, Jia Yang, Wenya Wang, Li Yang, and Rui Xia. 2024b. Generative multimodal data augmentation for low-resource multimodal named entity recognition. In *ACM Multimedia 2024*.

Jintao Liu, Chenglong Liu, and Kaiwen Wei. 2024. Multi-view prompt for fine-grained multimodal named entity recognition and grounding. In *ECAI 2024*, pages 2693–2700. IOS Press.

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*, pages 459–474, Cham. Springer International Publishing.

Christos Louizos and Max Welling. 2016. Structured and efficient variational deep learning with matrix gaussian posteriors. In *International conference on machine learning*, pages 1708–1716. PMLR.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018a. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018b. Visual attention model for name tagging in multimodal social media. In *ACL*, pages 1990–1999.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Hyunjong Ok, Taeho Kil, Sukmin Seo, and Jaeho Lee. 2024. Scanner: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. *arXiv preprint arXiv:2404.01914*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

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

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2024a. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*.

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

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023b. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 2391–2399, New York, NY, USA. Association for Computing Machinery.

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*, pages 5925–5936.

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*, pages 3176–3189.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7322–7329.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2024. Small models are valuable plug-ins for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 283–294.

Linyi Yang, Shuibai Zhang, Zhuohao Yu, Guangsheng Bao, Yidong Wang, Jindong Wang, Ruochen Xu, Wei Ye, Xing Xie, Weizhu Chen, et al. 2024. Supervised knowledge makes large language models better in-context learners. In *ICLR*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL*, 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.

Jianxing Yu, Qinliang Su, Xiaojun Quan, and Jian Yin. 2021. Multi-hop reasoning question generation and its application. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):725–740.

Jianxing Yu, Shiqi Wang, Han Yin, Qi Chen, Wei Liu, Yanghui Rao, and Qinliang Su. 2025. Diversified generation of commonsense reasoning questions. *Expert Systems with Applications*, 263:125776.

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 *AAAI*, 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 *AAAI*, volume 32.

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 Web Conference 2024*, pages 4047–4058.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A   Preliminary of Model's Uncertainty Estimation.

Even though deep learning networks have achieved remarkable success across various domains, they inevitably produce errors. Thus, measuring their uncertainty has become a crucial research direction, reflecting the reliability and trustworthiness of the model (Hu et al., 2023). In machine learning, uncertainty can be categorized into aleatoric uncertainty and epistemic uncertainty (Xiao and Wang, 2019; Gawlikowski et al., 2023). The former refers to the inherent uncertainty in the data due to its noise or randomness. The latter represents the model's uncertainty, which stems from a lack of knowledge or capabilities within the model itself (Hu et al., 2023).

Conventional deep learning models commonly employ a softmax layer for classification. An intuitive method for estimating model's uncertainty is to use 1 minus the probability outcomes from the softmax layer. However, the softmax operation produces deterministic point estimates that often yield overconfident for misclassified predictions (Guo et al., 2017). A common approach to calibrate model uncertainty is Bayesian Neural Network (BNN) (Lakshminarayanan et al., 2017; Kendall and Gal, 2017). In Bayesian Neural Networks (BNNs), the model parameters ψ are variables that follow a specific distribution. Given a

labeled dataset $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, BNNs fit a posterior probability distribution $p(\psi|\mathcal{S})$. For a classification model, the distribution of predictions $\mathbf{y}^*$ based on an input $\mathbf{x}^*$ can be expressed as:

$$p(\mathbf{y}^* = \phi_c|\mathbf{x}^*, \mathcal{S}) = \int \underbrace{p(\mathbf{y}^* = \phi_c|\mathbf{x}^*, \psi)}_{\text{Data}} \underbrace{p(\psi|\mathcal{S})}_{\text{Model}} d\psi$$

(10)

However, $p(\psi|\mathcal{S})$ is usually intractable based on Bayesian Posterior's rules. To address this, Variational Inferences (VI) (Blundell et al., 2015; Louizos and Welling, 2016) are often employed to approximate the Bayesian posterior by optimizing a tractable distribution, denoted as $p(\psi|\mathcal{S}) \approx q(\psi)$. Among them, a common approach is Monte Carlo Dropout (MCD) (Gal and Ghahramani, 2016), which involves randomly dropping neurons during the inference phase. This technique results in different predictions for each dropout configuration. By performing T different sampling, we obtain T distributions, which can be used to approximate the Bayesian posterior as:

$$P(\mathbf{y}^*|\mathbf{x}^*, \mathcal{S}) \approx \frac{1}{T} \sum_{i=1}^T P(\mathbf{y}^*|\mathbf{x}^*, \hat{\psi}^{(i)}),$$

$$\hat{\psi}^{(i)} \sim q(\psi), \quad i = 1, ..., T$$

(11)

In this way, by applying dropout to each layer, we can simply transform any deep neural network into a Bayesian neural network while maintaining the same training objectives as non-Bayesian networks. Unlike non-Bayesian methods, MCD activates dropout during inference. The model's uncertainty is quantified through variations in probability or entropy values (Zhang et al., 2024) across multiple sampling.

B   Details of Experiment Settings

B.1   Implementation Details

Small Model of UnCo. The experiments of small fine-tuned model is implemented on one NVIDIA RTX3090 GPU with Pytorch 1.9.1. For a fair comparison with baselines, we use the pre-trained BART-base model[5] as our backbone, ViT-B/32 from pre-training CLIP[6] as the visual encoder, and VinVL[7] as a class-agnostic RPN. During training, we set the batch size to 32, the learning rate to 3×10^{-5}, and the training epoch to 30. Our model

---

[5]https://huggingface.co/facebook/bart-base
[6]https://huggingface.co/openai/clip-vit-base-patch32
[7]https://github.com/pzzhang/VinVL

uses an AdamW optimizer and the number of candidate regions is set to 28. The hyperparameter γ and λ are selected in Appendix C.4. To mitigate the complexity of uncertainty estimation, we set the beam search size of BART as 1. The sampling times for Monte Carlo dropout are set to 10. Three uncertainty thresholds are determined as mentioned in Section 3.4. For Twitter-FMNERG, the span, region, and type uncertainty thresholds are set to 0.3, 0.4, 0.2, respectively.

MLLMs of UnCo. In this study, we conduct experiments on four different MLLMs, including Qwen2.5VL-7B, InternVL3 9B, Qwen2.5VL-72B, GPT4o, and Gemini-2.5 Pro. For Qwen2.5VL-7B[8] and InternVL3-9B[9], the experiments are implemented on two NVIDIA RTXA6000 GPUs. For Qwen2.5VL-72B[10], GPT4o, and Gemini-2.5 Pro, the experiments are based on official APIs. The temperature parameter of MLLM is set to 0.1. The details of prompts are shown in Figure 7, 8, and 9.

B.2 Baselines Systems

To evaluate the performance of UnCo, we select various baseline systems for comparison, categorized into three main types: pipeline methods, end-to-end frameworks, and Multimodal Large Language Models (MLLMs).

For the pipeline methods, we employ different state-of-the-art Multimodal Named Entity Recognition (MNER) models to extract text entities and then utilize object detectors (OD), including VinVL (Zhang et al., 2021b) or Faster R-CNN (Girshick, 2015), to identify visual entities. Finally, text and visual entities are matched by an Entity-aware Visual Grounding (EVG) (Yu et al., 2023) module.

For the MLLM approaches, we use 3-shot in-context learning with prompt-based chain-of-though (Wei et al., 2022) as comparative baseline (see the details in Figure 10). Furthermore, we proposed a *Direct Correction* pipeline as an extra comparison. This pipeline applies the same small model as UnCo and then directly feeds its results to a large model for correction.

These methods are described in detail as follows:

- **GVATT-OD-EVG** (Lu et al., 2018a) employs a visual attention mechanism integrated with a BiLSTM-CRF framework to extract multimodal entities.

- **UMT-OD-EVG** (Yu et al., 2020) introduces a multimodal transformer designed to capture cross-modality semantics effectively.

- **UMGF-OD-EVG** (Zhang et al., 2021a) addresses text-image integration through a multimodal graph fusion approach.

- **ITA-OD-EVG** (Wang et al., 2022b) leverages image-text translation and object tags to explicitly align visual and textual features.

- **MMT5/BARTMNER-OD-EVG** (Yu et al., 2023) enhances generative models T5/BART with a cross-modal transformer layer.

- **H-Index** (Yu et al., 2023) formulates the GM-NER task as sequence generation using a multimodal BART model with a pointer mechanism.

- **TIGER** (Wang et al., 2023a) is a T5-based generative model that transforms all span-type-region triples into target paraphrase sequences.

- **GDMA** (Li et al., 2024b) extends existing GMNER model with multimodal data augmentation using InstructBLIP and Stable Diffusion.

- **MQSPN** (Tang et al., 2025) is a query-based framework that aligns multimodal entities with learnable queries and generates them through set prediction.

- **GEM** (Wang et al., 2024b) is a knowledge argumentation framework that utilizes ChatGPT's knowledge to enhance fine-grained textual entities, fine-tuning two multimodal large language models (LLaVA and BLIP2) to generate top-k visual regions, with open-set SAM further employed for accurate visual entity grounding.

- **RiVEG** (Li et al., 2024a) is a pipeline framework for GMNER, which introduces a visual Entailment and entity expansion expressions to address weak image-text correlation and the gap between named entities and referring expressions, achieving state-of-the-art performance across all GMNER subtasks.

- **Qwen2.5VL** (Bai et al., 2025) is the latest flagship model of the Qwen multimodal large

language model, featuring enhanced visual recognition, precise object localization, and robust structured data extraction from multimodal content, with world knowledge up to 2024.

- **InternVL3** (Zhu et al., 2025) is the latest flagship model of InternVL, showcasing superior multimodal perception and reasoning capabilities across various tasks, achieving state-of-the-art zero-shot visual grounding performance in RefCoCo, RefCoCo+, and RefCoCog.

- **GPT4o** (Hurst et al., 2024) is the latest multimodal large language model from OpenAI, designed for real-time understanding and generation across text, vision, and audio modalities, with strong reasoning and interactive capabilities.

- **Gemini 2.5 Pro** (Comanici et al., 2025) is the advanced multimodal foundation model of Google DeepMind's Gemini series, integrating cutting-edge language understanding, visual reasoning, and tool-use abilities, achieving state-of-the-art performance across diverse benchmarks.

B.3 GMNER Datasets

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

Table 4: The statistics of two GMNER datasets.

In this study, we explore two tweet datasets: Twitter-GMNER (Yu et al., 2023) and Twitter-FMNERG (Wang et al., 2023a). Twitter-GMNER consists of four entity types—Person (PER), Organization (ORG), Location (LOC), and Others (OTHER)—for text-image pairs. Twitter-FMNERG builds on GMNER by incorporating 8 coarse-grained and 51 fine-grained entity types. These datasets are based on two publicly available MNER Twitter datasets, Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018b). Statistical details for Twitter-GMNER and Twitter-FMNERG are presented in Table 4.

B.4 Evaluation Metrics

Evaluation Metrics of GMNER. The GMNER prediction consists of entity span, type, and visual region. Following previous research (Yu et al., 2023), the correctness of each prediction is calculated as follows:

$$C_e/C_t = \begin{cases} 1, & \text{if } p_e/p_t = g_e/g_t; \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

$$C_r = \begin{cases} 1, & \text{if } p_r = g_r = \text{None}; \\ 1, & \text{if } \max(\text{IoU}_1, \dots, \text{IoU}_j) > 0.5; \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

where C_e, C_t, and C_r denote the correctness of entity span, type, and region predictions, respectively; p_e, p_t, and p_r are the predicted entity span, type, and region; g_e, g_t, and g_r are the ground truth span, type, and region; IoU_j is the Intersection over Union score between p_r and the j-th ground truth bounding box $g_{r,j}$.

The GMNER task uses precision (Pre.), recall (Rec.), and F1 score as evaluation metrics:

$$correct = \begin{cases} 1, & \text{if } C_e \text{ and } C_t \text{ and } C_o; \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

$$
\begin{aligned}
Pre &= \frac{\#correct}{\#predict}, \\
Rec &= \frac{\#correct}{\#gold}, \\
F1 &= \frac{2 \times Pre \times Rec}{Pre + Rec}
\end{aligned} \quad (15)
$$

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

Evaluation Metrics for Uncertainty Estimation. Following (Zhang et al., 2024; Hu et al., 2023), we utilize the Area Under the Receiver-Operator Characteristic Curve (AUROC) to evaluate the performance of binary classification between correct and incorrect predictions. The formulation is as follows:

$$AUROC = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[u(t_0) < u(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|} \quad (16)$$

where $\mathcal{D}^0$ is the set of correct predictions, and $\mathcal{D}^1$ is the set of incorrect predictions. $\mathbf{1}[u(t_0) < u(t_1)]$ denotes an indicator function which returns 1 if

Settings	Twitter-GMNER			Twitter-FMNERG		
	GMNER	MNER	EEG	GMNER	MNER	EEG
UnCo	**62.14**	**81.33**	**66.72**	**52.44**	**67.21**	**65.68**
w/o CoT	61.62	81.04	66.41	51.73	66.79	65.27
w/o UNC	61.75	80.98	66.56	51.50	66.64	65.08

Table 5: Ablation studies on Uncertainty-aware Hierarchical Correction.

$u(t_0) < u(t_1)$ otherwise return 0. In this paper, $u(t)$ is the uncertainty estimation function. Higher AUROC values indicate better performance.

C Additional Experimental Results

C.1 More Ablation Studies on Uncertainty-aware Hierarchical Correction

To further investigate the effectiveness of components within the Uncertainty-aware Hierarchical Correction module, we conducted additional ablation experiments on the corrector prompt. Specifically: (1) **w/o CoT**: We remove the prompt-based chain-of-thought, making the model directly output corrected results. (2) **w/o UNC**: We eliminated uncertainty and its contextual description from the corresponding prompt. The experimental results indicate that the performance of UnCo decreases with the removal of each component. Removing CoT results in performance drops of 0.52% and 0.71% on the Twitter-GMNER and Twitter-FMNERG datasets, respectively. Eliminating UNC leads to decreases of 0.39% and 0.94%. This demonstrates that both CoT and uncertainty effectively guide MLLMs in refining pre-detected entities, thereby enhancing overall performance.

C.2 Analysis of MLLM Correction

Correction	Twitter-GMNER	Twitter-FMNERG
Span Correction	12.32%	16.81%
Region Correction	28.03%	30.29%
Type Correction	15.74%	26.67%

Table 6: The proportion of MLLM Correction.

We analyzed the proportion of corrections performed by different correctors across various datasets, as shown in Table 6. The results reveal that the Span Corrector is invoked the least frequently. This is primarily because entity spans often exhibit strong domain specificity, which is largely determined by annotation guidelines. In contrast, region corrections occur more frequently

compared to both span and type corrections. This indicates a higher proportion of out-of-domain visual regions in the dataset, aligning with our initial motivation. For instance, even for the same "person" entity, variations in context, such as differing scenes or clothing styles, often make it challenging for small models to recognize them accurately. Additionally, in the Twitter-FMNERG dataset, we observed a higher frequency of type corrections. This is mainly due to the dataset's fine-grained entity types, which lead to underfitting for certain low-resource types. By leveraging MLLMs to further refine these entities, we can effectively enhance the performance of GMNER. However, as uncertain entities increase, the frequency of invoking MLLMs also rises, leading to higher computational resources. Uncertainty threshold is a trade-off solution to control the number of MLLM calling.

C.3 Density Distribution of Twitter-FMNERG

Figure 6 shows the density distributions and AUROC values for correct and incorrect predictions in Twitter-FMNERG. Compared to confidence scores, uncertainties estimated by MCD (MRD) have better performance to filter out incorrect predictions.

C.4 Hyperparameter Selection of UnCo

Table 7 shows the hyperparameter tuning experiments on UnCo's baseline model. We select $\gamma = 0.6, \lambda = 0.5$ for Twitter-GMNER dataset, and $\gamma = 0.6, \lambda = 0.3$ for Twitter-FMNERG dataset.

D The Prompt of MLLM Corrector

Figure 7, 8, and 9 are the designed prompts for Span Corrector, Type Corrector, Region Corrector, respectively. Figure 10 is the prompt used for MLLM baseline. Figure 11 shows a comprehensive correction process of UnCo.

Metric (F1%)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Twitter-GMNER										
γ	56.85	56.98	56.96	57.04	57.13	**57.21**	57.19	56.78	56.82	56.67
λ	56.89	56.80	56.76	56.92	**57.14**	57.07	56.85	56.71	56.63	56.58
Twitter-FMNERG										
γ	46.84	46.90	46.82	47.05	47.19	**47.28**	47.23	47.01	46.85	46.76
λ	46.81	46.94	**47.09**	47.03	46.98	46.86	46.95	46.72	46.69	46.52

Table 7: Results for UnCo (small) under different hyperparameters .
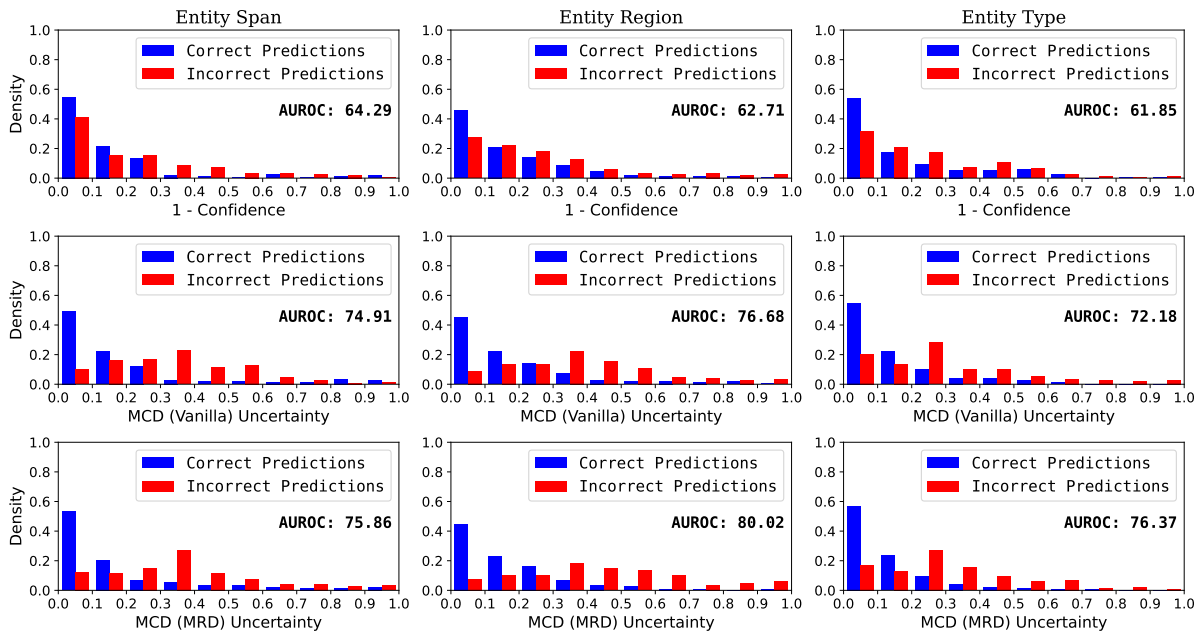
Figure 6: The density distribution of correct and incorrect predictions under different uncertainty intervals on Twitter-FMNERG. The uncertainty values are normalized to [0,1]. Higher AUROC values indicate better filtering of incorrect samples. MCD (Vanilla) uses only MC Dropout in UnCo, while MCD (MRD) includes additional training for modality representation debiasing.

System prompt

[Role]:
    You are an AI assistant focused on correcting named entity.
[Format Output Description]:
    Eg. ```json{"reasoning_process": "....", "corrected_entity": "...."}```
[Action Required]:
    - Carefully review the pre-detect textual named entity with uncertainty.
    - Please think step by step:
        1. What is the background knowledge of "entity" according to the original text?
        2. Is the span of the pre-detected entity correct? If not, what should the correct span be?
    - If you think the span of the entity is inaccurate, please correct the boundary of its span. Otherwise, just output the original prediction.
    - Important note: When you are correcting the span of an entity, please focus on tiny boundaries modification (one or two words) around the span and do not have additional outputs.
    - Uncertainty Description: Uncertainty refers to the confidence level of the pre-extraction results; higher uncertainty (>[threshold]) indicates a greater likelihood of errors. Please review carefully.
    - Please output your reasoning process and your final Corrected entity according to [Format_output_Description].

User

<Original Text>: 📖
<Pre-detected Entity> : ......
<Pre-detected Entity Uncertainty>: ......

MLLM-Response (Json Format):

```json
{
    "reasoning_process": "....",
    "corrected_entity": "...."
}
```

Figure 7: The prompt details designed for Span Corrector.

System prompt

[Role]:
    You are an AI assistant focused on correcting the type of the named entity.
[Format Output Description]:
    Eg. ```json{"reasoning_process": "....", "corrected_type": "...."}```
[Entity Type Description] :
    ......
    (The content of this part is determined by the specific dataset.)
    ......
[Action Required]:
    - Carefully review the pre-detect textual named entity and its type with uncertainty.
    - Please think step by step:
        1. What is the background knowledge of "entity" according to the original text and image?
        2. Is the type of the pre-detected entity correct?
    - If you think the type of the pre-detected entity is correct, no modification is needed. If you think its type is incorrect, please refer to [Entity Type Description] to correct it. Make decision based on textual context, visual cues and background knowledge.
    - Uncertainty Description: Uncertainty refers to the confidence level of the pre-extraction results; higher uncertainty (>[threshold]) indicates a greater likelihood of errors. Please review carefully.
    - Please output your reasoning process and your final Corrected entity type according to [Format_output_Description].

User
<Original Text>: 📖
<Image>: 📈
<Pre-detected Entity> : ......
<Pre-detected Entity type> : ......
<Pre-detected Entity type Uncertainty>: ......

MLLM-Response (Json Format):

```json
{
    "reasoning process": "....",
    "corrected_type": "...."
}
```

Figure 8: The prompt details designed for Type Corrector.

> **System prompt**

[Role]:
    You are an AI assistant focused on correcting the bounding box of the named entity from the provided image.
[Format Output Description]:
    Eg. ```json{"reasoning_process": "....", "corrected_bounding_box": "...."}```
[Action Required]:
    - Carefully review the provided image and the pre-detected entity, its type, the bounding box, as well as the corresponding uncertainty of predictions.
    - Please think step by step:
      1. What is the background knowledge of "entity" according to the original text and image?
      2. Is the bounding box provided by the pre-detected entity accurate?
    - If you think the bounding box provided by the pre-detected entity is correct, no modification is needed. If you think its bounding box is incorrect, please correct it. Some correct regions are not explicitly aligned to named entities, please think based on background knowledge and the visual cues within the image.
    - Important note: If the entity cannot be precisely located at a specific position within the image, or if the entity encompasses the entire image area, set the bounding box to 'Null'. If you are also uncertainty about its groundings, please output 'Null'.
    - Uncertainty Description: Uncertainty refers to the confidence level of the pre-extraction results; higher uncertainty (>[threshold]) indicates a greater likelihood of errors. Please review carefully.
    - Please output your reasoning process and your final Corrected bounding box according to [Format_output_Description].

> **User**

<Original Text>: 📖
<Image>: 🖼
<Pre-detected Entity> : ......
<Pre-detected Entity type> : ......
<Pre-detected Bounding box>: ......
<Pre-detected Bounding box Uncertainty>: ......

> **MLLM-Response (Json Format):**
```json
{
     "reasoning process": "....",
     "corrected_bounding_box": "...."
}
```

Figure 9: The prompt details designed for Region Corrector.

> **System prompt**

[Role]:
    You are an AI assistant focused on extracting the multimodal named entity from the provided image.
[Format Output Description]:
    Eg. ```json{"pre_entities": [{"phrase": "James", "entity_type": "PER", "region_box": [293, 21, 593, 449]}]}```
[Entity Type Description] :
    ......
     (The content of this part is determined by the specific dataset.)
    ......
[Action Required]:
    - Carefully review the provided image and the original text.
    - Please think Step-by-step about the following question:
      1. What entities are there in the original text?
      2. What are the types of these entities?
      3. What are the region-boxes of these entities?
    - Important note: If the entity cannot be precisely located at a specific position within the image, or if the entity encompasses the entire image area, set the bounding box to '[]'. Only output one nearest entity type.
   - Please output your final results according to [Format_output_Description].

> **User**

Here are some examples you can refer to:
[Input]: 🖼 📖   [Output]: 📖
[Input]: 🖼 📖   [Output]: 📖
[Input]: 🖼 📖   [Output]: 📖

[Input]:
   <Original Text>: 📖
   <Image>: 🖼

> **MLLM-Response (Json Format):**
```json
{
     "pre_entities": [
         {
             "phrase": "....",
             "entity_type": "....",
             "region_box": [...., ...., ...., ....]
         }
     ]
}
```

Figure 10: The prompt used in MLLM baseline, which has 3-shot demonstrations for in-context learning.

User

Text: RTO@ josh hafner :RT @ dmpl: Des Moines Tribune 11/22/1963. # JFK50 # desmoines @jasonnobleDMR

Image: Des Moines Tribune — Kennedy Is Slain By Texas Sniper

UnCo (Small)

Entity 1
Span: josh hafner, $u = 0.12$
Type: ORG, $u = 0.73$
Region: [], $u = 0.41$

Entity 2
Span: Moines Tribune, $u = 0.67$
Type: LOC, $u = 0.81$
Region: [], $u = 0.74$

Regions:

Note: No regions are output by UnCo(Small).

UnCo (Large)

- In the context of the original text and the image, 'josh hafner' refers to a writer and producer for USA today's nation now team... This is evident from the newspaper on the image that a signature appears in the end of an article. The pre-detected entity type is 'ORG', which is uncertain due to error. Carefully review both the image and context, a correction is needed. According to 'josh hafner' mentioned above, it should be corrected as 'PER'.

Type Corrector

- The pre-detected text "Moines Tribune" has an uncertainty score of 0.67, indicating high confidence. It looks like the entity refers to the "Des Moines Tribune", a former newspaper based in Des Moines, Iowa... The span is incomplete, as it omits the word "Des." The text "Des Moines Tribune" should replace the pre-detected entity.

Span Corrector

- The text "Des Moines Tribune" refers to the name of a newspaper, as seen clearly in the image. The original type "LOC" has an uncertainty score of 0.81, which is high, suggesting a significant likelihood of error. Based on the image and context, it is a proper noun and represents a publication rather than a physical location or geographic region. Therefore, the entity type "LOC" (Location) initially assigned to it is incorrect. The correct entity type should be "MISC" (Miscellaneous), as it pertains to a media organization.

Type Corrector

- The text "Des Moines Tribune" refers to the name of a newspaper. It is a media organization The pre-detected bounding box is listed as [], meaning no bounding box was initially assigned. However, the pre-detected region has an uncertainty score of 0.74, which is moderately high. Carefully review the image, I find "Des Moines Tribune" is prominently displayed at the top center of the newspaper page, indicating that it is the title of the publication. Therefore, the output bounding box should capture the area in the image where the text "Des Moines Tribune" is located. The corrected bounding box is [94, 55, 408, 87].

Region Corrector

Figure 11: A comprehensive process of span corrector, type corrector and region corrector in UnCo.